Проблемы прикладной физики

Обработка изображений и цифровая фильтрация

Под редакцией
Т. Хуанга

Издательство „Мир"
Москва 1979

Обработка изображений и цифровая фильтрация

Topics in Applied Physics
Founded by Helmut K. V. Lotsch
Volume 6

Picture Processing and Digital Filtering

Edited by T. S. Huang

With Contributions by
H. C. Andrews, F. C. Billingsley,
J. G. Fiasconaro, B. R. Frieden, T. S. Huang,
R. R. Read, J. L. Shanks, S. Treitel

Springer-Verlag
Berlin Heidelberg New York 1975

Проблемы прикладной физики

Обработка изображений и цифровая фильтрация

Под редакцией Т. Хуанга

Перевод с английского
канд. техн. наук Е. З. Сороки и
канд. техн. наук В. А. Хлебородова

Издательство «Мир»
Москва 1979

УДК 621.391.837 : 681.3

Книга посвящена актуальной проблеме цифровой обработки двумерных изображений, к которым относятся медицинские и промышленные рентгенограммы, электронно-микроскопические снимки, радиолокационные и звуколокационные изображения, карты сейсмической активности, телевизионные изображения и снимки, получаемые со спутников. Цифровая обработка производится с целью генерирования и воспроизведения сообщений, улучшения их качества, извлечения информации, распознавания образов, эффективного кодирования для передачи и запоминания изображений и т. п. Приводятся достаточно полные и современные сведения по таким важным вопросам цифровой обработки, как двумерные преобразования, нерекурсивная и рекурсивная двумерная фильтрация, улучшение и реставрация изображений и учет аппаратурных шумов.

Книга может служить справочным руководством для исследователей и инженеров и учебным пособием по цифровой обработке изображений.

Редакция литературы по физике

2704070200

X $\frac{20405,30401-070}{041(01)-79}$ 70—79

© by Springer-Verlag Berlin Heidelberg 1975. All rights reserved. Authorized translation from English language edition published by Springer-Verlag Berlin — Heidelberg — New York

© Перевод на русский язык, «Мир», 1979

От переводчиков

Цифровая обработка изображений стала в настоящее время важнейшей и вполне самостоятельной областью применения ЦВМ. В широком смысле она охватывает улучшение изображений, их кодирование, распознавание образов и машинную графику. Предлагаемая книга посвящена обработке изображений в более узком аспекте, включающем только улучшение и кодирование.

Поток журнальных статей по этим вопросам велик и продолжает нарастать, тогда как количество книжных публикаций явно недостаточно (из последних зарубежных публикаций следует отметить [1—3]). Книги, вышедшие на русском языке [4—9], лишь в какой-то степени удовлетворяют имеющуюся потребность в классификации, обобщении и оценке многочисленных методов обработки. Ценность данной книги в том и состоит, что она была задумана именно как своеобразный путеводитель в области обработки изображений. Книга охватывает следующие вопросы: исследование двумерных преобразований, проектирование двумерных рекурсивных и нерекурсивных фильтров [10], рассмотрение методов улучшения и реставрации изображений, анализ влияния аппаратурного шума [11]. Подробные характеристики основных пяти глав книги содержатся в специальной вводной главе, написанной проф. Хуангом, которая ориентирует читателя в представленном материале.

Книгу отличает широта охвата, трезвая оценка достигнутых результатов и возможностей и высокий уровень изложения. Пожалуй, несколько необычным для такой «математизированной» книги является особое внимание, уделяемое авторами чисто практическим вопросам, а именно влиянию шума на результат обработки и учету требуемого времени вычислений.

Можно надеяться, что книга окажется полезной не только для специалистов в области обработки изображений с помощью универсальных ЦВМ, но и для тех, кто занимается обработкой

изображений в реальном масштабе времени с использованием специализированных цифровых быстродействующих устройств (например в телевидении).

Перевод книги выполнили Е. З. Сорока (гл. 1, 2 и 5) и В. А. Хлебородов (предисловие, гл. 3, 4 и 6).

Литература

1. *Ahmed N., Rao K. R.*, Orthogonal Transforms for Digital Signal Processing. — New York: Springer-Verlag, 1975.
2. *Pratt W. K.*, Digital Image Processing. — New York: J. Wiley, 1977.
3. *Rosenfeld A., Kak A. C.*, Digital Picture Processing. — New York: Academic Press, 1977.
4. Иконика: Сб. статей/Отв. ред. Д. С. Лебедев. — М.: Наука, выпуски 1968, 1970, 1975 г.; Вопросы кибернетики, вып. 38, 1978 г.
5. *Ярославский Л. П.*, Устройства ввода-вывода изображений для цифровых вычислительных машин. — М.: Энергия, 1968.
6. *Розенфельд А.* Распознавание и обработка изображений с помощью вычислительных машин. Пер. с англ. под ред. Д. С. Лебедева. — М.: Мир, 1972.
7. *Эндрюс Г.* Применение вычислительных машин для обработки изображений. — М.: Энергия, 1977.
8. Обработка изображений при помощи цифровых вычислительных машин. Под ред. Г. Эндрюса и Л. Инло. Пер. с англ. под ред. Д. С. Лебедева. — М.: Мир, 1973.
9. *Ярославский Л. П., Мерзляков Н. С.*, Методы цифровой голографии. — М.: Наука, 1977.
10. *Рабинер Л., Гоулд С.*, Теория и применение цифровой обработки сигналов. Пер. с англ. — М.: Мир, 1978.
11. *Васьков С. Т., Мамонтов Г. М, Поташников А. К., Ткач С. Е.* Сканирующие устройства на ЭЛТ высокого разрешения.—Новосибирск: Наука, 1978.

Предисловие

С обработкой сигналов сталкиваются при решении каждой научной или технической задачи. Многие сигналы многомерны, т. е. они являются функциями нескольких переменных. В качестве примеров можно привести рентгеновские снимки в медицине и промышленности, электронно-микроскопические снимки, радиолокационные и звуколокационные карты, сейсмические данные, телевизионные изображения и снимки со спутников (подобных спутнику ERTS [1])). Цели обработки могут быть различными: генерирование и отображение данных, повышение качества, извлечение информации, распознавание образов, эффективное кодирование при передаче или запоминании данных и т. д. Примерами весьма полезных и очень интересных применений обработки многомерных сигналов могут служить: распознавание знаков, улучшение снимков Луны и Марса, полученных со спутников, составление карт земных ресурсов по снимкам со спутника ERTS и трансаксиальная томография.

Обработка сигналов может быть цифровой или аналоговой. Конечно, цифровые методы намного превосходят аналоговые по гибкости. Многие задачи обработки многомерных сигналов удалось решить в последнее время только благодаря быстрому развитию цифровой технологии.

Данная книга знакомит читателя с углубленным рассмотрением избранных проблем цифровой обработки двумерных сигналов (т. е. изображений): гл. 2—4 посвящены двумерным преобразованиям и фильтрам, гл. 5 освещает вопросы улучшения и реставрации изображений, а гл. 6 содержит рассмотрение влияния шума аппаратуры цифровой обработки (в основном сканеров). Все перечисленные главы задуманы как ознакомительные,

---

[1]) Спутник, предназначенный для отработки технологии дистанционного зондирования земных ресурсов. — *Прим. перев.*

тем не менее они подводят читателя к самым передовым рубежам проводимых исследований. Авторам представляется, что эта книга окажется полезной как справочник для ученых и инженеров или как дополнительный учебный материал в курсах по цифровой обработке сигналов, обработке изображений и цифровой фильтрации.

Авторы выражают благодарность ARPA [1]) за финансирование издания этой книги и работы, описанной в вводной первой главе (контракт MDA 703-74-0098). Авторы признательны проф. Ломану, предложившему идею о необходимости такой книги.

Зап. Лафейетт, шт. Индиана
Апрель 1975 г.

Т. Хуанг

[1]) Управление перспективного планирования научно-исследовательских работ США. — *Прим. перев.*

1. Введение

Т. Хуанг

1.1. Что такое обработка изображений?

Обработка изображений в широком смысле слова означает выполнение различных операций над многомерными сигналами (т. е. сигналами, которые являются функциями многих переменных). К многомерным сигналам относятся, например, телевизионные изображения, фотографии разведывательного характера, медицинские рентгенограммы, электронно-микроскопические фотографии молекул, радио- и звуколокационные карты, диаграммы сейсмических данных.

Цели, преследуемые при обработке таких многомерных сигналов, весьма различны. Однако в большинстве случаев они могут быть отнесены к одной из следующих четырех категорий: улучшение изображений, эффективное кодирование изображений, распознавание образов и машинная графика.

Во многих случаях целью обработки изображений является улучшение их качества. Ярким примером в этой области могут служить работы по улучшению изображений Марса и Луны, проведенные Лабораторией реактивного движения Калифорнийского технологического института. Публикации в газетах позволили широкой публике увидеть изображения до и после обработки. Было достигнуто просто поразительное улучшение обработанных изображений по сравнению с необработанными. Если считать, что обработка производится для устранения внесенных ранее искажений, то такую обработку называют реставрацией изображений.

Когда для передачи или запоминания сигнала требуется канал с чрезмерной пропускной способностью или память слишком большой емкости, желательно использовать более эффективное кодирование и тем самым снизить требования к каналу и памяти. В качестве примера можно привести видеотелефон. При обычных методах передачи канал, требуемый для видеотелефона, по пропускной способности в несколько сотен раз превышает телефонный канал. Без эффективного кодирования видеотелефон вряд ли окажется жизнеспособным с экономической точки зрения. Другим примером может служить проблема хранения рентгенограмм в больницах. При огромном количестве рентгенограмм хранение их требует много места, а проблема поиска

необходимой рентгенограммы оказывается почти неразрешимой. Выход из этого положения возможен лишь с помощью эффективного кодирования.

Понятие «распознавание образов» мы толкуем довольно свободно, имея в виду обнаружение образов (или другой информации) и их выделение из сигналов. При таком подходе это понятие охватывает широкий спектр методов — от простого подсчета среднего значения сигнала до распознавания Ньютоном природного «образа», который мы называем законом тяготения. В области распознавания зрительных образов наибольшее внимание привлекло направление, задачей которого является распознавание знаков. Выпускается много моделей машин для распознавания печатных знаков, и особенно для распознавания знаков определенного шрифта. Однако распознавание рукописных знаков все еще остается трудно решаемой задачей. В последнее время стало довольно активно развиваться еще одно направление — применение автоматического распознавания образов в качестве средства медицинской диагностики. Уже начинают появляться работоспособные системы такого назначения. Следует упомянуть еще одно развивающееся направление — автоматическое составление карт земных ресурсов на основе фотографий, полученных с помощью спутников.

Машинная графика — это область обработки изображений, которая охватывает ввод изображений в ЭВМ и их вывод из ЭВМ, соответствующее программирование и проблемы, связанные с организацией данных. Интересным направлением машинной графики является воспроизведение трехмерных объектов. Возможно, что путем использования голографических методов со временем удастся обеспечить действительно трехмерное воспроизведение.

1.2. Структура книги и библиография общего характера

В этой книге представлены результаты исследований по избранным вопросам цифровой обработки изображений. Особое внимание уделено реставрации изображений, тогда как эффективное кодирование обсуждается лишь вкратце (в гл. 2), а распознавание зрительных образов и машинная графика вообще не рассматриваются.

При цифровой обработке изображений используют различные математические методы, как линейные, так и нелинейные. Однако в общем и систематизированном виде могут быть рассмотрены только линейные методы (в частности, унитарные преобразования и линейные операции, инвариантные относительно сдвига, т. е. пространственно-инвариантные операции). Полез-

ной для практических целей общей теории нелинейных операций не существует, поэтому, когда такие операции встречаются в конкретных приложениях, каждую из них приходится рассматривать в отдельности.

Главы 2—4 посвящены линейным методам, применяемым при обработке изображений. В них рассматриваются соответственно двумерные преобразования, нерекурсивные и рекурсивные фильтры. В гл. 5 обсуждается реставрация изображений, включая многие нелинейные методы обработки. Наконец, в гл. 6 рассматриваются аппаратурные проблемы обработки изображений и особенно проблемы, связанные с шумом.

Читателям, не знакомым с обработкой изображений и желающим получить общее представление о предмете, можно рекомендовать книги Розенфельда [1] и Хуанга и др. [2]. Полезно также просмотреть два специальных выпуска журнала ТИИЭР, один из которых посвящен цифровой обработке изображений [3], а другой — цифровому распознаванию образов [4]. Для читателей, интересующихся более глубокими сведениями по конкретным разделам обработки изображений, наиболее полезными будут обзорные статьи Розенфельда [5—7].

1.3. Двумерные преобразования

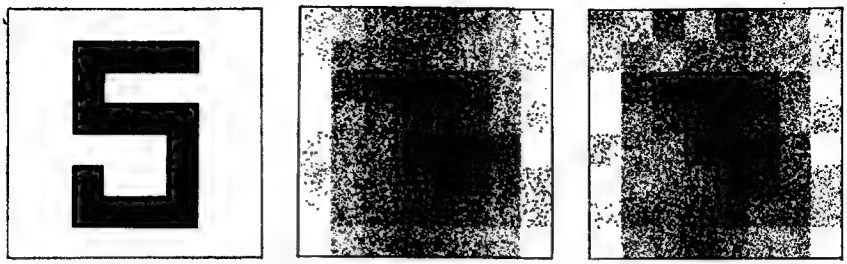
Остальная часть данного введения посвящена пояснениям к последующим главам книги. Некоторые из этих пояснений, возможно, станут более понятными после прочтения соответствующих глав.

В гл. 2 Эндрюс рассматривает двумерные преобразования в обобщенной форме разложений по внешним произведениям. Основные сведения о преобразованиях и быстрых алгоритмах их вычисления читатель может найти в книге Эндрюса [8]. Наиболее интересную часть гл. 2 составляют, по-видимому, приложения так называемого разложения по сингулярным значениям (РСЗ) к обработке и, в частности, к реставрации изображений. Как показано в этой книге, РСЗ тесно связано с псевдоинверсией матриц. Глубокий анализ данного предмета содержится в книге Ланцоша [9].

При реставрации изображений применение РСЗ особенно эффективно для борьбы с шумом, присутствующим во всех искаженных изображениях. Используя обозначения гл. 2, представим неразделимое линейное пространственно-инвариантное искажение в виде

$$\boldsymbol{g} = [H]\boldsymbol{f} + \boldsymbol{n}, \tag{1.1}$$

где $\boldsymbol{f}$ и $\boldsymbol{g}$ — матрицы-столбцы, содержащие отсчеты исходного объекта и искаженного изображения соответственно. Количество элементов в $\boldsymbol{g}$ и $\boldsymbol{f}$ не обязательно должно быть одинаковым.



а *б* *в*

Фиг. 1.1. *а* — оригинал; *б* — размытое изображение с аддитивным гауссовым шумом (среднее значение 0, среднеквадратичное отклонение 0,1); *в* — размытое изображение с аддитивным гауссовым шумом (среднее значение 0, среднеквадратичное отклонение 0,5).

Прямоугольная матрица $[H]$ находится из импульсной характеристики искажающей системы; $\boldsymbol{n}$ — матрица-столбец, содержащая отсчеты шума. Шум может образовываться, например, в результате действия детектора излучения. Задача состоит в следующем: даны $\boldsymbol{g}$ и $[H]$, оценить $\boldsymbol{f}$.

Хорошей оценкой является

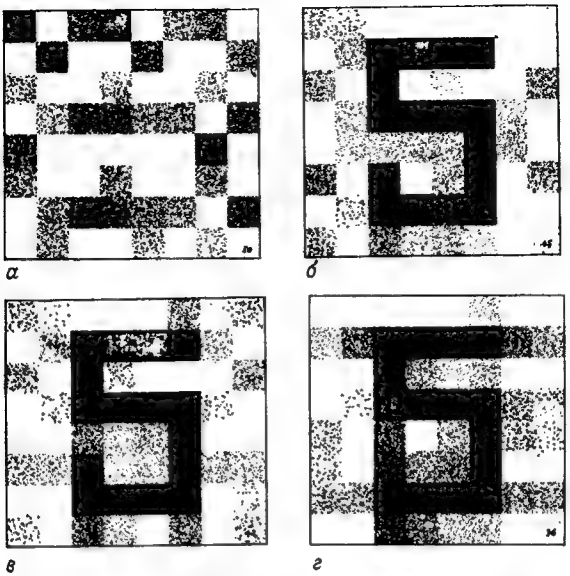
$$\hat{\boldsymbol{f}} = [H]^{+}\boldsymbol{g}, \tag{1.2}$$

где $[H]^{+}$ — псевдоинверсия $[H]$ по Муру — Пенроузу. Приятное свойство псевдоинверсии состоит в том, что она всегда существует, так что нам не нужно беспокоиться о том, имеет ли система линейных уравнений (1.1) решение и является ли оно единственным. Фактически $\hat{\boldsymbol{f}}$ представляет собой минимальное по норме решение системы (1.1), полученное по методу наименьших квадратов при $\boldsymbol{n} = 0$. В присутствии шума имеем

$$\hat{\boldsymbol{f}} = [H]^{+}[H]\boldsymbol{f} + [H]^{+}\boldsymbol{n}, \tag{1.3}$$

где первый член в правой части представляет собой минимальную по норме оценку, полученную по методу наименьших квадратов в отсутствие шума, а второй член — добавку, вносимую шумом. К сожалению, во многих случаях действие шума оказывается преобладающим, и сигнальная часть (1.3) может быть полностью замаскирована.

Применение РСЗ для вычисления псевдоинверсии позволяет полностью исправить это положение. Как указано в гл. 2, имеется возможность построить обобщенный винеровский фильтр. Однако для демонстрации эффективности РСЗ достаточно просто заметить, что ценой увеличения шума можно улучшить качество

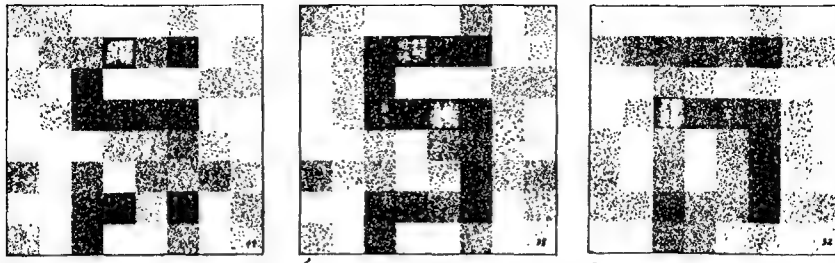


Фиг. 1.2. Изображение, полученное реставрацией фиг. 1.1, *б* с использованием РСЗ. Для изображений *а*, *б*, *в* и *г* использовано соответственно 50, 48, 42 и 36 членов.

сигнала, выбирая число членов, используемых при РСЗ псевдоинверсии. Используя РСЗ $[H]^+$, преобразуем (1.3) к виду

$$\hat{f} = \sum_{i=1}^{R} \lambda_i^{-1/2} \boldsymbol{v}_i \boldsymbol{u}_i^t \{[H]f\} + \sum_{i=1}^{R} \lambda_i^{-1/2} \boldsymbol{v}_i \boldsymbol{u}_i^t \boldsymbol{n}, \tag{1.4}$$

где $\boldsymbol{u}_i$ и $\boldsymbol{v}_i$ — собственные векторы $[H][H]^t$ и $[H]^t[H]$ соответственно, а λ_i — собственные значения. В общем члены первой суммы имеют более или менее сравнимые значения, в то время как значения членов второй суммы растут пропорционально $1/\sqrt{\lambda_i}$ (λ_i расположены в порядке убывания амплитуд). С увеличением числа используемых членов в суммах, входящих в (1.4), первая сумма все сильнее приближается к исходному объекту, а отношение сигнал/шум (отношение первой суммы ко второй) становится все меньшим. Желательно получить разумное соотношение между этими двумя эффектами. Одна из возможностей состоит в том, чтобы остановиться на члене, при котором значение шума становится сравнимым со значением сигнала. Лучший вариант получается, если следить за результатом суммирования после добавления каждого нового члена и остановить процедуру по достижении визуально наилучшей реставрации.



а *б* *в*

Фиг. 1.3. Изображение, полученное реставрацией фиг. 1.1, *в* с использованием РСЗ. Для изображений *а*, *б* и *в* использовано соответственно 44, 38 и 32 члена.

На фиг. 1.1—1.3 показаны примеры реставрации, полученные путем моделирования на ЦВМ. Исходным изображением служила цифра 5, дискретизированная на 8×8 точек. Матрица размера 8×8, представляющая квантованный исходный объект, показана на фиг. 1.1, *а*. Каждой точке внутри цифры придано значение 7, а каждой точке вне ее — значение 0. Это изображение было смазано путем замены каждой его точки средним значением девяти точек, образующих блок размером 3×3, сцентрованный относительно рассматриваемой точки. После этого к изображению был добавлен случайный гауссов шум. Два искаженных изображения показаны на фиг. 1.1, *б* и *в*, причем среднеквадратичное отклонение шума в этих изображениях составляет 0,1 и 0,5 соответственно.

Реставрация выполнялась с использованием выражения

$$\hat{f} = \sum_{i=1}^{P} \lambda_i^{-1/2} v_i u_i^t g. \tag{1.5}$$

Для каждого искаженного изображения испытывались значения $P = 1, 2, \ldots, 64$ и контролировались все 64 реставрации. Некоторые из отобранных реставрированных изображений показаны на фиг. 1.2 и 1.3. Эти изображения воспроизводились на электростатическом плоттере с модуляцией плотности точек. Каждая точка изображения размера 8×8 представлялась квадратным блоком. Оценка 64 собственных значений $[H]\,[H]^t$ показала, что 15 из них практически не отличаются от нуля (они гораздо меньше остальных значений). Отсюда с очевидностью следует, что при вычислении (1.5) не требуется использовать более 49 членов. Это иллюстрирует фиг. 1.2, *а*. Следует, однако, отметить, что при сильном шуме может оказаться желательным использование еще меньшего количества членов. Так, для иска-

женного изображения на фиг. 1.1, *в* визуально наилучшую реставрацию дают, по-видимому, 38 членов, как видно из фиг. 1.3.

Мы убедились, что метод вычисления псевдоинверсии с помощью РСЗ вполне пригоден для реставрации изображений, подвергнутых действию шума и линейных искажений. Однако этому методу присущ один существенный недостаток: даже при изображениях среднего размера требуется находить собственные векторы и собственные значения очень больших матриц. Например, для изображения 100×100 точек получится матрица $10\,000 \times 10\,000$. В гл. 2 показано, что если искажающая импульсная характеристика разделима, то задачу можно существенно упростить. Однако интересно, что́ можно сделать, когда импульсная характеристика неразделима. Одна из возможностей, о которой говорится в гл. 2, состоит в аппроксимации неразделимой импульсной характеристики суммой разделимых характеристик. Другой подход заключается в следующем. Пространственная протяженность искажающей импульсной характеристики обычно гораздо меньше, чем протяженность изображения. Поэтому размер матрицы можно сократить путем разбиения искаженного изображения на фрагменты меньшего размера (которые остаются все же гораздо большими, чем искажающая импульсная характеристика) и реставрировать каждый фрагмент в отдельности. При этом возникает проблема, связанная с краевым эффектом. Дело в том, что точки каждого фрагмента, лежащие вблизи границы, зависят от точек соседних фрагментов, так что теоретически нельзя считать один фрагмент независимым от других. Математическое решение этой проблемы в настоящее время не известно. Однако в одной работе [10] указывается, что можно воспользоваться методом Винера — Хопфа. Практически, чтобы использовать метод разбиения, следует образовать перекрывающиеся фрагменты, реставрировать каждый из них в отдельности, а затем исключить границы.

1.4. Двумерные цифровые фильтры

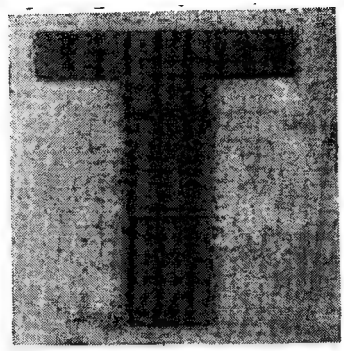
Двумерные (линейные пространственно-инвариантные) цифровые фильтры применяются во многих областях обработки изображений, таких, как улучшение изображений, реставрация изображений, подвергнутых линейным искажениям, обнаружение образов с помощью согласованной фильтрации и выделение контуров. В гл. 3 (написанной Фиасконаро) и гл. 4 (написанной Ридом и др.) даются обзоры методов проектирования двумерных нерекурсивных и рекурсивных цифровых фильтров.

Хотя методы проектирования одномерных цифровых фильтров разработаны относительно хорошо [11, 12], их обобщение

на два измерения сопряжено с трудностями. В случае нерекурсивных фильтров неприятности связаны просто с количеством измерений. В случае же рекурсивных фильтров возникает дополнительная трудность: полином от двух переменных в общем случае не может быть разложен в произведение двух сомножителей первого порядка. Поэтому проверка устойчивости в двух измерениях становится чрезвычайно сложной. Кроме того, это означает, что двумерный рекурсивный фильтр общего вида не может быть реализован в виде комбинации фильтров низкого порядка, что позволило бы ослабить эффекты квантования и шум округления. Справедливости ради следует, пожалуй, указать, что в настоящее время не существует хороших универсальных методов проектирования двумерных рекурсивных фильтров. Заметим, что применение рекурсивных фильтров оправдывается тем, что потенциально они требуют меньшего времени вычисления, чем нерекурсивные фильтры. Поэтому трудности, связанные с проектированием рекурсивного фильтра, стоит преодолевать только при условии, что эта потенциальная возможность может быть реализована.

В течение последних лет исследования в области двумерных цифровых фильтров сильно активизировались. Большая часть этих исследований была посвящена рекурсивным фильтрам. Поэтому в отличие от гл. 3, охватывающей почти все известные методы проектирования двумерных нерекурсивных фильтров, в гл. 4 не вошли некоторые из самых последних работ в области проектирования рекурсивных фильтров. В этой связи мы хотим обратить внимание читателей на работы Мариа и Фахми [13] и Сид-Ахмеда и Жюльена [14].

Эти авторы исходят из положения, что для обеспечения свободной манипуляции эффектами квантования и ошибками округления рекурсивные фильтры необходимо синтезировать в виде комбинации звеньев низкого порядка. Поэтому они используют каскады звеньев второго порядка и определяют коэффициенты фильтров методами математической оптимизации. Для экономии времени вычислений были использованы методы минимизации без учета ограничений. В работе Мариа и Фахми проверка устойчивости производилась после каждой итерации, и если фильтр оказывался неустойчивым, то для обеспечения устойчивости величина шага уменьшалась. В работе Сид-Ахмеда и Жюльена устойчивость обеспечивалась путем соответствующего изменения переменных. Поскольку рациональная функция двух переменных в общем случае не может быть выражена в виде произведения сомножителей второго порядка, такой подход позволяет спроектировать лишь субоптимальные фильтры. Однако такие фильтры могут оказаться достаточно хорошими для практического применения.



Фиг. 1.4. Изображение, смазанное по горизонтали, с аддитивным шумом.

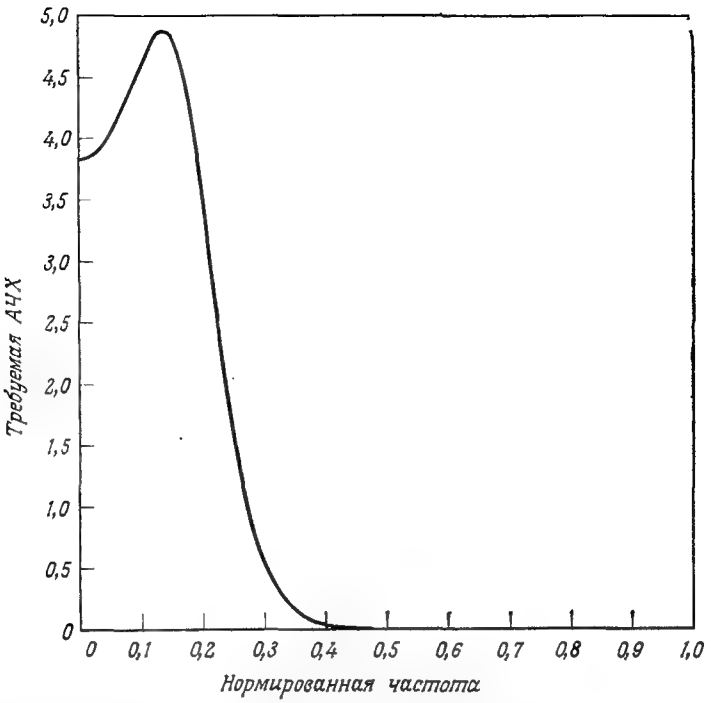
Прежде чем закончить обсуждение двумерных цифровых фильтров, нам хотелось бы особо отметить два взаимосвязанных и часто остающихся незамеченными обстоятельства.

1. Фаза преобразования Фурье обычно более важна, чем его амплитуда. Действительно, производя обратное преобразование Фурье спектра амплитуд (считая при этом фазы нулевыми), мы получаем пятно, не имеющее сходства с оригиналом. В то же время, производя обратное преобразование спектра фаз (считая спектр амплитуд постоянным), мы можем получить некоторое подобие исходного изображения.

2. Во многих приложениях к обработке изображений (таких, как согласованная фильтрация и реставрация изображений, искаженных в результате линейного перемещения с переменной скоростью) требуются фильтры с нелинейной фазочастотной характеристикой.

Из второго обстоятельства следует, что требуется проводить больше исследований, направленных на разработку методов проектирования, позволяющих определять как амплитудно-частотные характеристики (АЧХ), так и фазочастотные характеристики (ФЧХ). В частности, представляется полезным изучить возможности развития методов проектирования как одномерных, так и двумерных нерекурсивных фильтров с помощью линейного программирования, чтобы они позволяли учитывать требования к ФЧХ.

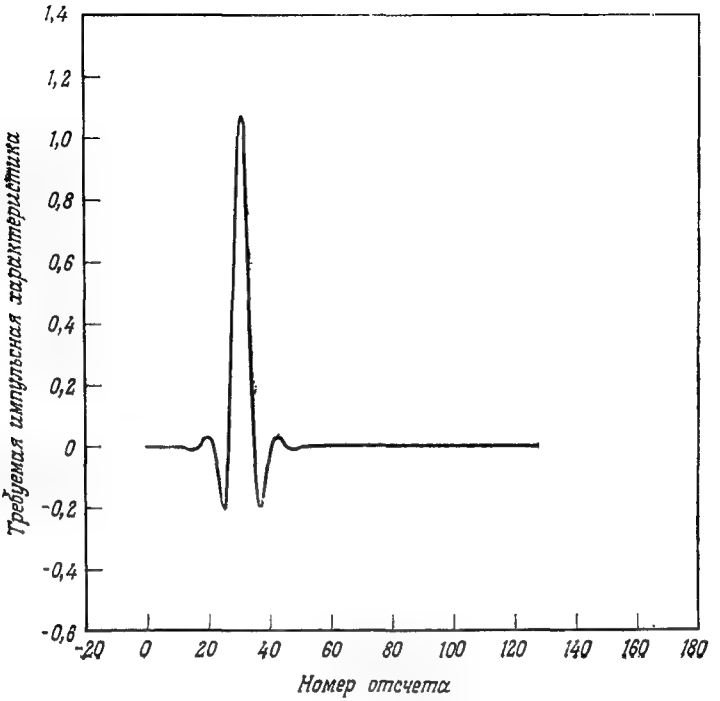
Из первого обстоятельства следует, что даже в случае проектирования рекурсивных фильтров для обработки изображений с частотной характеристикой, имеющей линейную ФЧХ, ее необходимо все-таки включать в требования и не считать, что если установлены требования к АЧХ, то можно надеяться на получение почти линейной ФЧХ, хотя именно таким образом обычно



Фиг. 1.5. Амплитудно-частотная характеристика среднеквадратичного инверсного фильтра.

поступают при проектировании одномерных рекурсивных фильтров. Когда имеется возможность обращения ко всему фильтруемому изображению, можно добиться получения действительно нулевой ФЧХ путем двукратного применения одного и того же рекурсивного фильтра, но начиная обработку с противоположных углов изображения. Однако такое решение оказывается неприемлемым, если фильтрацию требуется выполнить в реальном масштабе времени.

Чтобы проиллюстрировать значение первого из указанных выше обстоятельств, рассмотрим следующий простой пример. Изображение буквы Т дискретизировалось растром 128×128 точек. Каждой точке внутри буквы придавалось значение 5, а каждой точке вне ее — значение 0. Это изображение подвергали линейному смазыванию в горизонтальном направлении с использованием гауссовой весовой функции со стандартным отклонением 3 точки. Затем в изображении был введен белый гауссов шум со стандартным отклонением 0,25, так что в результате

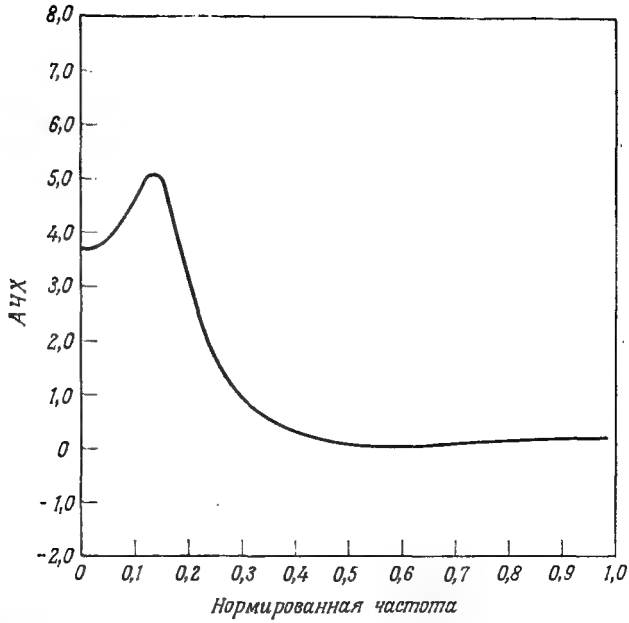


Фиг. 1.6. Импульсная характеристика среднеквадратичного инверсного фильтра.

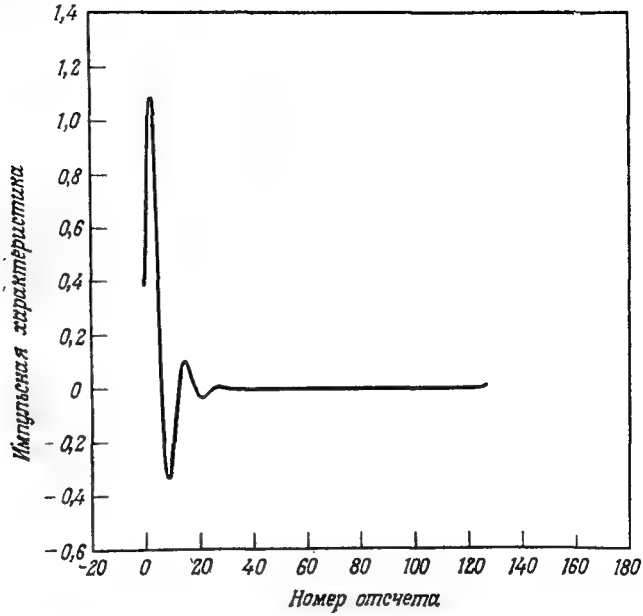
получилось смазанное и зашумленное изображение, показанное на фиг. 1.4.

Среднеквадратичный инверсный фильтр [15] для линейного гауссова смазывания был рассчитан с использованием представления спектра сигнала нарастающим косинусом. АЧХ этого инверсного фильтра показана на фиг. 1.5. ФЧХ имеет нулевое значение. Импульсная характеристика показана (с задержкой) на фиг. 1.6.

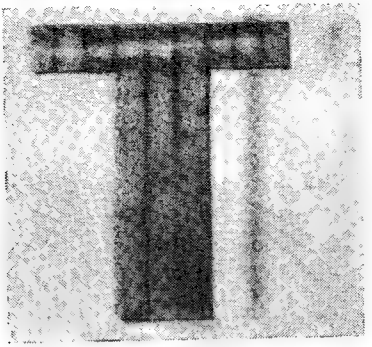
Для реставрации изображения с помощью рекурсивной фильтрации был спроектирован рекурсивный фильтр четвертого порядка (с 4 полюсами и 4 нулями), АЧХ которого аппроксимирует характеристику инверсного фильтра (фиг. 1.5). При этом был использован метод, аналогичный методу Дечки [16]. АЧХ рекурсивного фильтра, показанная на фиг. 1.7, почти тождественна соответствующей характеристике идеального инверсного фильтра (фиг. 1.5). Однако импульсная характеристика рекурсивного фильтра (фиг. 1.8) заметно отличается от импульсной



Фиг. 1.7. Амплитудно-частотная характеристика спроектированного рекурсивного фильтра.



Фиг. 1.8. Импульсная характеристика спроектированного рекурсивного фильтра.



Фиг. 1.9. Результат реставрации изображения на фиг. 1.4, выполненной с использованием спроектированного рекурсивного фильтра.

характеристики идеального инверсного фильтра (фиг. 1.6). Поэтому при реставрации с помощью такого рекурсивного фильтра изображения, показанного на фиг. 1.4, получается изображение с паразитными узорами (фиг. 1.9).

1.5. Улучшение и реставрация изображений

При реставрации изображений мы стремимся компенсировать искажения, вносимые в изображения в процессе их формирования системами отображения. К типичным причинам искажений относятся движение камеры, аберрации объектива, низкочастотный характер электрооптических систем и турбулентность атмосферы. При улучшении изображений наряду с компенсацией искажений или вместо нее мы стремимся придать изображению форму, более удобную для наблюдения человеком или для дальнейшей обработки на ЦВМ. В качестве примеров улучшения изображений укажем обострение контуров и получение цветных изображений из синтезированных составляющих.

В гл. 5, написанной Фриденом, представлен обзор методов реставрации изображений. В настоящем разделе мы приводим некоторые дополнительные соображения по следующим вопросам: итерационные методы; методы, основанные на математическом программировании; линейные пространственно-зависимые искажения; рекурсивное оценивание.

В гл. 5 обсуждаются итерационные методы Ван Циттерта — Джанссона. Здесь мы займемся другим итерационным методом — методом проекций [17], имеющим определенные преимущества. Представим монохромное изображение функцией двух переменных, причем обе независимые переменные являются пространственными координатами, а значение функции выражает

яркость изображения в данной точке. Тогда задача реставрации изображения может быть сформулирована математически следующим образом. Пусть дано искаженное изображение

$$g(x, y) = D[f(x, y)] + n(x, y), \tag{1.6}$$

где $f(x, y)$ — оригинал, D — искажающий оператор, а $n(x, y)$ — шум; требуется найти хорошую оценку $f(x, y)$.

Один общий подход к решению задачи реставрации изображения состоит в попытке решить (1.6) итерационными методами. При k-й итерации приближенное решение $f^{(k)}(x, y)$ определяется через приближенное решение при предыдущей итерации $f^{(k-1)}(x, y)$ и разность $g(x, y) - D[f^{(k-1)}(x, y)]$. В качестве хорошего начального приближения можно взять

$$f^{(0)}(x, y) = g(x, y). \tag{1.7}$$

Предположим, что оператор D линейный. Тогда, пренебрегая шумом, дискретизированный вариант (1.6) можно выразить в виде

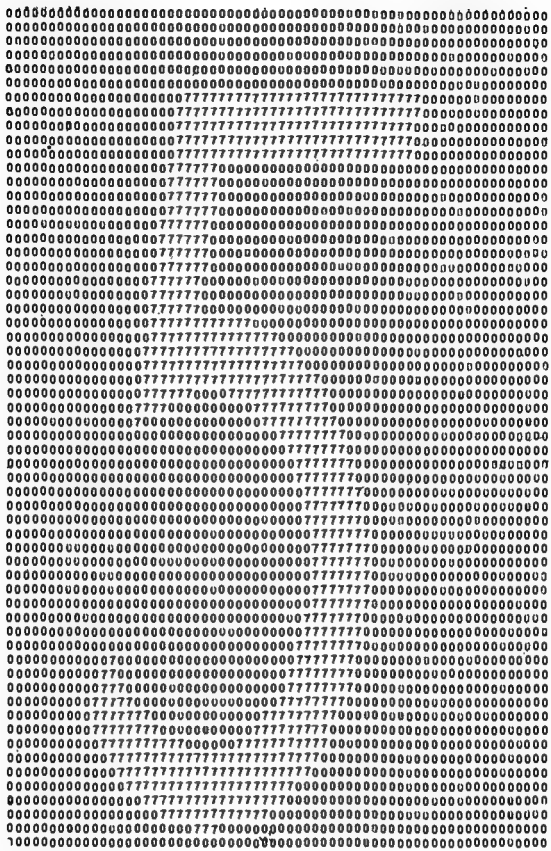
$$\begin{aligned} a_{11}f_1 + a_{12}f_2 + \ldots + a_{1N}f_N &= g_1, \\ a_{21}f_1 + a_{22}f_2 + \ldots + a_{2N}f_N &= g_2, \\ &\vdots \\ a_{M1}f_1 + a_{M2}f_2 + \ldots + a_{MN}f_N &= g_M, \end{aligned} \tag{1.8}$$

где f_i и g_i — отсчеты $f(x, y)$ и $g(x, y)$ соответственно, а коэффициенты a_{ij} — постоянные. Количество отсчетов функций f и g составляет N и M соответственно.

Для описания метода проекций удобнее использовать геометрический подход. Будем рассматривать $\mathbf{f} = (f_1, f_2, \ldots, f_N)$ как вектор или точку в N-мерном пространстве. Тогда каждое уравнение в (1.8) представляет гиперплоскость. Пусть начальное приближение будет $\mathbf{f}^{(0)} = (f_1^{(0)}, f_2^{(0)}, \ldots, f_N^{(0)})$. Следующее приближенное решение $\mathbf{f}^{(1)}$ представляет собой проекцию $\mathbf{f}^{(0)}$ на гиперплоскость $a_{11}f_1 + \ldots + a_{1N}f_N = g_1$, т. е.

$$\mathbf{f}^{(1)} = \mathbf{f}^{(0)} - \frac{(\mathbf{f}^{(0)} \cdot \mathbf{a}_1 - g_1)}{\mathbf{a}_1 \cdot \mathbf{a}_1} \mathbf{a}_1, \tag{1.9}$$

где $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1N})$, а точка $(\cdot)$ обозначает обычное скалярное произведение. Затем найдем проекцию $\mathbf{f}^{(1)}$ на гиперплоскость $a_{21}f_1 + \ldots + a_{2N}f_N = g_2$ и назовем ее $\mathbf{f}^{(2)}$ и т. д., пока не найдем $\mathbf{f}^{(M)}$, удовлетворяющую последнему уравнению в (1.8). На этом заканчивается первый цикл итераций. После этого начинаем снова с первого уравнения системы (1.8): находим проекцию $\mathbf{f}^{(M)}$ на $a_{11}f_1 + \ldots + a_{1N}f_N = g_1$ и называем ее $\mathbf{f}^{(M+1)}$, затем



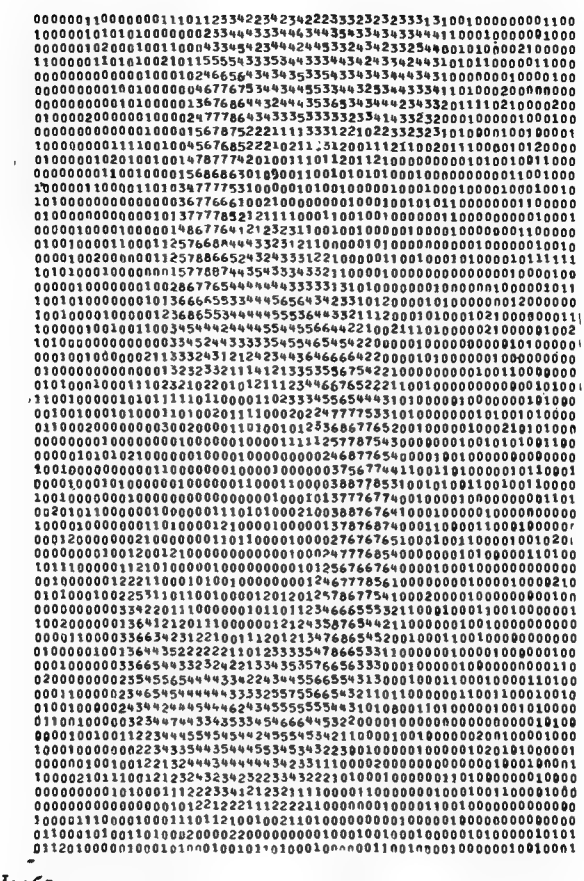
Фиг. 1.10. Исходное изображение.

находим проекцию $f^{(M+1)}$ на $a_{21}f_1 + \ldots + a_{2N}f_N = g_2$ и т. д., пока не дойдем до последнего уравнения в (1.8) и тем самым не завершим второй цикл итераций. Можно показать, что при продолжении процесса итераций векторная последовательность $f^{(0)}$, $f^{(M)}$, $f^{(2M)}$, $f^{(3M)}$, ... всегда сходится для любых заданных N, M и a_{ij}, так что

$$\lim_{k\to\infty} f^{(kM)} = f. \tag{1.10}$$

Можно показать, что если система уравнений (1.8) имеет единственное решение, то этим решением будет f, а если эта система имеет бесконечное множество решений, то f будет решением, которое минимизирует

$$|f - f^{(0)}|^2 = (f_1 - f_1^{(0)})^2 + (f_2 - f_2^{(0)})^2 + \ldots + (f_N - f_N^{(0)})^2. \tag{1.11}$$

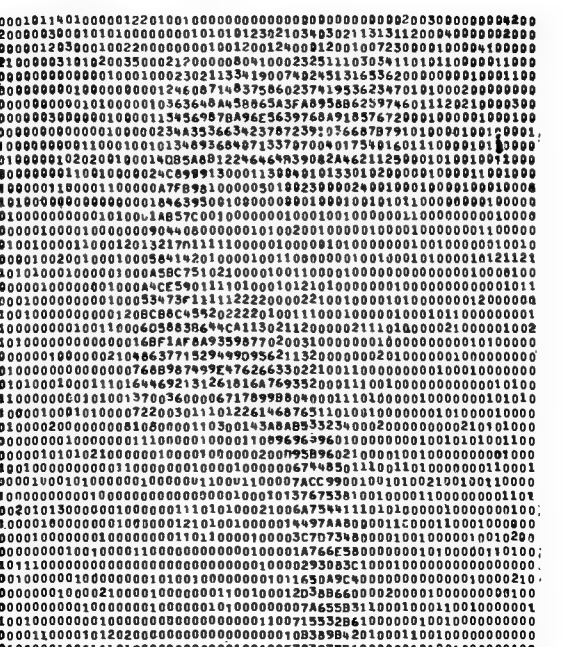


Фиг. 1.11. Изображение, смазанное по вертикали, с аддитивным шумом.

Следовательно, при бесконечном множестве решений мы можем надеяться на получение хорошего решения, если начнем с хорошего приближения $f^{(0)}$.

При применении метода проекций можно использовать различную априорную информацию об изображении. В частности, ограничение $f_i \geqslant 0$ можно использовать, приравнивая любые отрицательные составляющие $f^{(k)}$ нулю перед нахождением следующей проекции. Если же известно, что исходное изображение ограничено определенной областью, то в (1.8) можно приравнять нулю те a_{ij}, которые соответствуют отсчетам f_i, лежащим вне этой области.

Некоторые результаты моделирования на ЦВМ представлены на фиг. 1.10—1.12. Изображение цифры 5 дискретизировалось



Фиг. 1.12. Результат реставрации изображения на фиг. 1.11, полученный с использованием метода проекций (после пяти циклов итерации).

растром 64×64 точки. Каждой точке, принадлежащей цифре, придано значение 7, а каждой точке вне ее — значение 0. Это изображение было подвергнуто линейному смазыванию, и для реставрации оригинала в присутствии шума использовался метод проекций.

В изображениях, показанных на фиг. 1.10—1.12, каждая точка квантовалась на 16 уровней (со значениями от 0 до 15), причем уровни от 10 до 15 обозначались буквами от *A* до *F*. Исходное изображение, показанное на фиг. 1.10, было смазано по вертикали на 10 точек. Затем в смазанное изображение был введен шум, так что в результате было получено зашумленное и смазанное изображение, показанное на фиг. 1.11. Шум был белым,

независимым от сигнала и имел равномерное распределение со стандартным отклонением 0,5. К этому зашумленному и смазанному изображению был применен метод проекций. Результат, полученный после пяти итераций, показан на фиг. 1.12. Мы замечаем, что даже при таком относительно низком отношении сигнал/шум метод проекций дает удивительно хороший результат. Здесь в качестве начального приближения было принято $f^{(0)} = 0$.

Хотя в приведенном примере искажение было пространственно-инвариантным и одномерным, метод проекций, очевидно, можно применять для исправления любых линейных искажений (в общем случае пространственно-зависимых и двумерных). Если изображение содержит $N = n \times n$ точек, а импульсная характеристика искажающего оператора — самое большее $L = l \times l$ точек, то один цикл итерации требует выполнения приблизительно $2LN$ умножений и такого же количества сложений. Например, если изображение содержит $N = 200 \times 200$ точек, а искажающая импульсная характеристика содержит не более $L = 10 \times 10$ точек, то один цикл итерации потребует приблизительно $8 \cdot 10^6$ умножений и столько же сложений. Если ЦВМ выполняет одно умножение и одно сложение за 1 мкс, то один цикл итерации потребует около 8 с.

Другой подход к решению уравнений (1.8), когда g_i искажены шумом и имеется ограничение $f_i \geqslant 0$ (а также другие ограничения в виде равенств или неравенств), состоит в том, что задача формулируется на языке математического программирования. Например, можно пытаться искать такое $\mathbf{f}$, которое минимизирует

$$F(\mathbf{f}) = \sum_{i=1}^{M} \left(\sum_{j=1}^{N} a_{ij} f_j - g_i \right)^2 \tag{1.12}$$

при ограничениях

$$f_i \geqslant 0; \quad i = 1, 2, \ldots, N. \tag{1.13}$$

Но это типичная задача квадратичного программирования. Для детального ознакомления с методом читателю следует обратиться к работам Раста и Барраса [18] и Маскаренаса [19].

Метод Мак-Адама [20], упоминаемый в конце гл. 5, подобен (но не идентичен) методу математического программирования. Мак-Адам сделал попытку найти решения уравнений (1.8) с учетом ограничивающих неравенств для f_i, g_i и a_{ij}. В общем случае здесь существует бесконечное множество решений. Мак-Адам разработал эффективный с вычислительной точки зрения алго-

ритм для отыскания решения в случае линейного пространственно-инвариантного (ЛПИ) искажения.

Многие из методов, описываемых в гл. 5, равно как и рассмотренные выше методы проекций и математического программирования, а также обсуждаемый в гл. 2 метод разложения по сингулярным значениям (РСЗ) могут применяться для борьбы с линейными пространственно-зависимыми (ЛПЗ) искажениями общего вида. Однако все эти методы требуют весьма значительных вычислений. Для некоторых специфических видов ЛПЗ-искажений иногда удается найти более экономичные пути решения. Интересный пример представляют искажения, вызываемые аберрациями объективов. Оказалось, что, используя полярные координаты (r, θ) и производя соответствующую замену переменной r, аберрации типа комы можно свести к линейным пространственно-инвариантным искажениям [21]. Также с использованием полярных координат астигматизм и кривизна поля могут быть сведены к линейным пространственно-инвариантным искажениям в одном направлении и к ЛПЗ-искажениям в другом направлении. Благодаря этому существенно упрощается компенсация подобных искажений [22].

Проблему реставрации изображения можно рассматривать как проблему оценивания. Поэтому здесь могут применяться методы теории оценок. В последнее время некоторые авторы [23—25] использовали рекуррентные алгоритмы оценки (реализующие фильтры Калмана — Бьюси) для уменьшения шума в изображении. При этом изображение моделировалось двумерным марковским процессом. Такой подход легко обобщить для применения при реставрации изображений, в которые внесены линейные искажения и шум. Некоторые пути обобщения марковской модели были изучены Джейном и Анжелом [26].

В гл. 5 показано, а здесь мы также хотим это подчеркнуть, что хотя большинство прежних работ по реставрации изображений, искаженных ЛПИ-системами, посвящено применению инверсных ЛПИ-фильтров, такие инверсные фильтры оказались, по-видимому, не слишком эффективным средством получения реставраций хорошего качества. ЛПЗ- и нелинейные методы в общем гораздо более эффективны (особенно когда они позволяют учитывать априорную информацию об изображении). Однако большинство известных в настоящее время ЛПЗ- и нелинейных методов требуют весьма большого времени вычислений даже для обработки изображений средних размеров. Поэтому поиск методов этого класса, эффективных как по результатам, так и по времени вычислений, представляет собой одну из главных задач, которые должны решаться исследователями, работающими в этой области.

1.6. Технические средства цифровой обработки изображений

Решающим ограничивающим фактором в большинстве задач по обработке изображений является шум. В гл. 6 Биллингсли рассматривает проблему шума, уделяя основное внимание анализу шумовых свойств сканеров. Сканер преобразует аналоговое изображение в цифровые данные; он играет важную роль независимо от того, производится ли обработка на универсальной ЦВМ или на специализированном аппарате. В гл. 6 содержится много практически полезных результатов, значение которых для исследователей в области цифровой обработки изображений вряд ли можно переоценить.

Читателям следует учитывать происходящее в настоящее время быстрое развитие приборов на поверхностных акустических волнах и приборов с зарядовой связью, позволяющих выполнять одномерную свертку с весьма высокой скоростью [27]. Эти приборы могут также найти применение для ввода и обработки изображений.

Интересно отметить, что развитие технических средств оказывает сильное влияние на алгоритмы обработки. Например, чтобы воспользоваться приборами на поверхностных акустических волнах и с зарядовой связью для вычисления двумерного преобразования Фурье в реальном масштабе времени, было предложено свести его к преобразованию одномерных сверток [28].

1.7. Другие области обработки изображений

Хотя идеи и методы, рассматриваемые в этой книге, пригодны для использования во всех областях обработки изображений, основное внимание здесь уделяется улучшению и реставрации изображений. Сведения о других областях обработки изображений читатели могут почерпнуть из соответствующей литературы по эффективному кодированию изображений [29, 30], по распознаванию зрительных образов [31, 32] и по машинной графике [33, 34].

От автора. Данная работа финансировалась Управлением перспективного планирования научно-исследовательских работ США по контракту MDA 903-74-C-0098. Эксперименты по моделированию на ЭВМ были выполнены П. Нарендрой, Дж. Барнеттом и Д. Баркером.

Литература

1. *Rosenfeld A.*, Picture Processing by Computer, Academic Press, New York, 1969. (Имеется перевод: *Розенфельд А.* Распознавание и обработка изображений. — М.: Мир, 1972.)
2. *Huang T. S., Schreiber W. F., Tretiak O. J.*, Proc. IEEE, 59. 1588 (1971). [Имеется перевод: ТИИЭР, 59, № 11, 59 (1971)].
3. Proc. IEEE, Special Issue on Digital Picture Processing (July 1972). (Имеется перевод: ТИИЭР, 60, № 7 (1972), специальный выпуск «Обработка изображений при помощи цифровых вычислительных машин».)
4. Proc. IEEE, Special Issue on Digital Pattern Recognition (October 1972). (Имеется пеервод; ТИИЭР, 60, № 10 (1972), специальный выпуск «Распознавание образов при помощи цифровых вычислительных машин».)
5. *Rosenfeld A.*, Computing Survey, 5, 81 (1973).
6. *Rosenfeld A.*, Computer Graphics and Image Processing, 1, 394 (1972).
7. *Rosenfeld A.*, Computer Graphics and Image Processing, 3, 178 (1974).
8. *Andrews H.*, Computer Techniques in Image Processing, Academic Press, New York, 1970. (Имеется перевод: *Эндрюс Г.* Применение вычислительных машин для обработки изображений. — М.: Энергия, 1977.)
9. *Lanczos C.*, Linear Differential Operators, Van Nostrand, London, 1961, Chap. 3.
10. *Robbins G. M., Huang T. S.*, Inverse filtering for linearly shift-variant imaging systems, Proc. Symp. «Bildverarbeitung und interaktive Systeme», DLR-Mitt. 73—11, p. 40, (DFVLR-Forschungszentrum, Oberpfaffenhofen, West Germany, December 1971).
11. *Gold B., Rader C.*, Digital Processing of Signals, McGraw-Hill, New York, 1969. (Имеется перевод: *Голд Б., Рэйдер Ч.* Цифровая обработка сигналов. — М.: Советское радио, 1973.)
12. *Rabiner L., Rader C.* (eds.), Digital Signal Processing, IEEE Press, New York, 1972.
13. *Maria G. A., Fahmy M. M.*, IEEE Trans. on Acoustics, Speech, Signal Proc., ASSP-22, 15 (1974).
14. *Sid-Ahmed M. A., Julien G. A.*, Frequency domain design of a class of stable two-dimensional recursive filters (preprint), University of Windsor, Windsor, Ontario, Canada, June 1974.
15. *Helstrom C. W.*, J. Opt. Soc. Am., 57, 297 (1967).
16. *Deczky A. G.*, IEEE Trans. Audio Electroacoustics, AU-20, 257 (1972).
17. *Tanaba K.*, Numerical Mathematics, 17, 203 (1971).
18. *Rust B. W., Burrus W. R.*, Mathematical Programming and the Numerical Solution of Linear Equations, American Elsevier, New York, 1972.
19. *Mascarenhas N. D.*, Digital image restoration under a regression model — the unconstrained, linear equality and inequality constrained approaches, USCIPI Rep. 520, Image Processing Institute, University of Southern California, Los Angeles, Calif. (January 1974).
20. *McAdam D. P.*, J. Opt. Soc. Am., 60, 1617 (1970).
21. *Robbins G. M., Huang T. S.*, Proc. IEEE, 60, 862 (July 1972). [Имеется перевод: ТИИЭР, 60, № 7, 163—174 (1972).]
22. *Sawchuk A. A., Peyrovian M. J.*, Space-variant restoration of astigmatism and curvature of field, USCIPI Rep. 530, Image Processing Institute, University of Southern California, Los Angeles, Calif. (March 1974) p. 68.
23. *Habibi A.*, Proc. IEEE, 60, 878 (1972). [Имеется перевод: ТИИЭР, 60, № 7, 182 (1972).]
24. *Nahi N. E.*, Proc. IEEE, 60, 872 (1972). [Имеется перевод: ТИИЭР, 60, № 7, 174 (1972).]
25. *Kak A. C., Panda D.*, Statistical modeling and recursive filtering of images, Rep. TR-EE 74-22, Purdue University, West Lafayette, Indiana (June 1974), p. 35.

26. *Jain A. K., Angel E.*, IEEE Trans. Computers, C-32, 470 (1974).
27. *Huang T. S., Tretiak O. J.* (eds.), Proc. of Seminar on Signal Processing, NEREM, Boston, Mass. (1973).
28. *Speiser J. M., Whitehouse H. J.*, A two-dimensional discrete Fourier transform architecture, NUC TN 1221 Naval Undersea Center, San Diego, Calif. (October 1973).
29. *Huang T. S., Tretiak O. J.* (eds), Picture Bandwidth Compression, Gordon and Breach, New York, 1972.
30. *Huang T. S.*, Bandwidth Compression of Optical Images, Progress in Optics, Vol. 10, North-Holland, Amsterdam, 1972, p. 1.
31. *Duda R., Hart P.*, Pattern Classification and Scene Analysis, J. Wiley, New York, 1973. (Имеется перевод: *Дуда Р., Харт П.*, Распознавание образов и анализ сцен. — М.: Мир, 1976.)
32. *Cover T.*, IEEE Trans. Inf. Theory, IT-19, 827 (1973).
33. *Pankhurst P. J.*, в книге Advances in Information System Science, Vol. 3, ed. by J. T. Tou, Plenum Press, New York, 1970, pp. 215—282.
34. *Prince M. D.*, Interactive Graphics for Computer-Aided Design, Addison-Wesley, Reading, Mass., 1971.

2. Двумерные преобразования

Г. Эндрюс

Настоящая глава содержит общие сведения о двумерных преобразованиях и их применениях при цифровой обработке изображений на универсальных ЦВМ. Применена такая система обозначений, которая позволяет читателю, имеющему некоторое знакомство с матричной алгеброй, разобраться в основных принципах кодирования и реставрации изображений с использованием двумерных преобразований. Материал главы излагается в последовательности, соответствующей развертыванию математического представления предмета, и сопровождается примерами кодирования изображения, иллюстрирующими выполняемые преобразования как в исходной области (т. е. в области координат изображения), так и в преобразованной области (т. е. в области координат преобразования). Подробно рассматривается реставрация исходных изображений (оригиналов) с использованием численных методов матричной алгебры, позволяющих избежать получения вырожденных изображений. Внимательный читатель заметит, что ключевым моментом многих положений, формулируемых в этой главе, является предположение о разделимости. В главе содержится как вводный учебный материал, так и описания процедур реставрации, разработанных в самое последнее время, и автор надеется, что общая математическая основа, включающая двумерные преобразования и матричные квадратичные формы, позволит читателю постепенно перейти от материала, носящего в большей мере учебный характер, к результатам более новых исследований по обработке изображений.

2.1. Введение

2.1.1. Решаемые задачи — кодирование, реставрация, выделение признаков

Популярность цифровой обработки изображений резко возросла за последние несколько лет, причем особое развитие получили такие направления обработки, как кодирование, реставрация и выделение признаков изображений. Кодирование

изображений, как правило, связывают с попытками передать изображение по цифровой сети связи наиболее эффективным способом. Обычно при этом подразумевают изыскание возможности сократить число битов, описывающих изображение, и тем самым уменьшить число битов, которое требуется передавать по системе связи. Примерами таких систем могут служить национальные сети связи между ЦВМ и сети связи с временным уплотнением для проведения совещаний, обмена изображениями и проведения медицинских консультаций, системы передачи спектрозональных снимков со спутников, космических кораблей-зондов и аппаратов с дистанционным управлением, системы фототелеграфной передачи изображений от отпечатков пальцев до текста, а также системы передачи изображений по существующим телефонным сетям.

Если обратиться к истории, то кодирование изображений оказалось в центре внимания исследователей раньше, чем цифровая реставрация, и поэтому кодирование достигло, вероятно, более высокой стадии развития. Однако в последнее время интерес исследователей привлекли изображения, качество которых ухудшено по сравнению с оригиналами; таким образом, возникла потребность реставрировать изображения для получения улучшенной оценки оригиналов. Примерами подобных процедур могут служить исправление нарушенной фокусировки, устранение смазывания вследствие движения, реставрация искажений, вносимых системами отображения (в том числе коррекция аберраций, кривизны поля, перекоса плоскости пленки и других нарушений оптимальных условий отображения) и, наконец, фильтрация и подавление шума. Для методов реставрации характерно, что они обусловливаются либо имеющейся априорной информацией, либо моделью искажающего явления, действующего в данной системе отображения. Необходимость применения реставрации становится вполне очевидной, если учесть ту огромную роль, которую человек отводит зрению в обеспечении выживания. Учитывая определенные успехи, достигнутые при обработке одномерных сигналов (обычно имеющих аргументом время), можно ожидать значительно большего прогресса в области обработки двумерных визуальных сигналов.

В качестве возможных объектов цифровой обработки могут рассматриваться изображения, получаемые при космических и биомедицинских исследованиях, промышленные рентгенограммы, фотографии разведывательного характера, изображения, получаемые в телевизионных системах, в инфракрасных системах прямого обзора и в радиолокаторах бокового обзора, и, наконец, изображения, образуемые спектрозональными и некоторыми другими сложными системами отображения объектов и сцен в двумерный формат. Кроме того, объектом цифровой обра-

ботки являются многие виды синтетических изображений. К таким изображениям можно отнести различные данные, представляемые в двумерном формате для того, чтобы обеспечить более эффективное восприятие их человеком. Поэтому объектом приложения универсальных двумерных методов улучшения и реставрации могут стать также диаграммы дальность — скорость и дальность — время, «отпечатки» голоса, спектрограммы речи и т. п.

Восполняя семантический пробел, определим реставрацию как восстановление изображения, направленное на его приближение к исходному изображению (оригиналу) и осуществляемое путем инверсии определенного искажающего явления. Чтобы выполнить инверсию и тем самым осуществить реставрацию, необходимо располагать определенной информацией об искажающем явлении. Эта информация может существовать в виде аналитических или статистических моделей либо иных априорных сведений, основанных на знаниях (или предположениях) о физической системе, осуществляющей процесс отображения. Поэтому особое внимание следует уделять источникам искажений и описывающим их моделям.

Выделением признаков при цифровой обработке изображений называют выбор специфических характеристик, выражающих некоторые особенности изображения, для последующего распознавания образов, классификации, принятия решений и интерпретации. Полезными признаками могут быть фактура, форма, статистические параметры и т. п. В некоторых приложениях полезные признаки часто удается получить путем двумерного преобразования исходного изображения в пространство, где новые координаты оказываются статистически менее коррелированными, чем координаты исходного изображения. Затем в таком пространстве с менее выраженной корреляцией могут быть выделены признаки, более эффективные для целей классификации, поскольку некоррелированные характеристики требуют меньшей размерности для описания, чем коррелированные.

2.1.2. Линейные модели

Большинство моделей для анализа систем обработки изображений обычно аппроксимируется с использованием предположения об их линейности. Такая линейная модель отображения, по существу, предопределяет характер реставрации, необходимой для инверсии искажений, так же как и структуру системы передачи кодированного изображения (для случая линейной аппроксимации системы).

Для того чтобы обеспечить удобство и согласованность записей и облегчить понимание, введем следующие обозначения

(фиг. 2.1):

$f(\xi, \eta)$ — оригинал,

$g(x, y)$ — изображение,

$h(x, y, \xi, \eta)$ — импульсная характеристика, или функция рассеяния точки (ФРТ) в предположении, что система отображения линейна.

Множество различных методов, используемых для цифровой обработки изображений, обусловило появление ряда моделей, отражающих попытки реставрировать изображения путем эффективной инверсии искажающего явления, происходящего при отображении оригинала. Пусть $g(x, y)$ — изображение оригинала $f(\xi, \eta)$, искаженное оператором $h(x, y, \xi, \eta)$, так что

$$g(x, y) = \int_{-\infty}^{\infty}\int f(\xi, \eta)\, h(x, y, \xi, \eta)\, d\xi\, d\eta. \qquad (2.1)$$

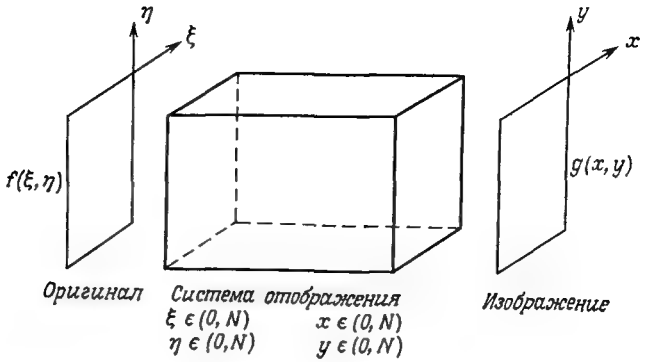
Искажение в системе описывается импульсной характеристикой, или функцией рассеяния точки h, которая физически соответствует выходу системы, когда на ее входе действует дельта-функция, или точечный источник света. Если при зондировании плоскости оригинала точечным источником вид импульсной характеристики сохраняется неизменным (исключая положение в плоскости изображения), то говорят, что система пространственно-инвариантна, т. е. имеется пространственно-инвариантная функция рассеяния точки (ПИФРТ) и действует соотношение

$$g(x, y) = \int_{-\infty}^{\infty}\int f(\xi, \eta)\, h(x - \xi, y - \eta)\, d\xi\, d\eta. \qquad (2.2)$$

В противном случае имеем систему с пространственно-зависимой функцией рассеяния точки (ПЗФРТ), и сохраняет силу выражение (2.1). Большинство исследователей удовлетворяются моделью (2.2) с вариациями, учитывающими аддитивный или мультипликативный шум и т. п. Для нахождения $f(\xi, \eta)$ по $g(x, y)$ вполне пригодны методы преобразования Фурье, которые позволяют определить инверсию h в (2.2), используя связь преобразования Фурье со сверткой. Так, в отсутствие шума

$$G_f(u, v) = H_f(u, v)\, F_f(u, v), \qquad (2.3)$$

где G_f, H_f и F_f — фурье-образы функций g, h и f соответственно, причем для определения $F_f(u, v)$ требуется лишь произвести инверсию H_f, если она существует. При использовании моделей с ПЗФРТ методы преобразования Фурье оказываются неприменимыми, и для реставрации приходится привлекать более универсальные и более мощные методы инверсии.



Фиг. 2.1. Модель линейной системы отображения.

Для инверсии обеих описанных выше систем часто применяются три основных линейных метода, которые можно назвать

а) непрерывно-непрерывным,

б) непрерывно-дискретным и

в) дискретно-дискретным.

При первом методе анализа весь процесс реставрации изображения рассматривается как непрерывный (хотя в итоге он неизбежно осуществляется как дискретный). При втором методе анализа принимается, что оригинал непрерывен, а изображение представлено в виде отсчетов и, следовательно, дискретно. При третьем методе предполагается, что все компоненты преобразования дискретны; для реставрации в нем используются только численный анализ и принципы линейной алгебры. Модели отображения имеют следующие аналитические выражения для каждого из трех названных методов:

а) непрерывно-непрерывная модель

$$g(x, y) = \int_{-\infty}^{\infty}\int f(\xi, \eta)\, h(x, y, \xi, \eta)\, d\xi\, d\eta, \tag{2.4}$$

б) непрерывно-дискретная модель

$$g_i = \int_{-\infty}^{\infty}\int h_i(\xi, \eta) f(\xi, \eta)\, d\xi\, d\eta \quad i = 1, \ldots, N^2, \tag{2.5}$$

в) дискретно-дискретная модель

$$\boldsymbol{g} = [H]\boldsymbol{f}. \tag{2.6}$$

В непрерывно-непрерывной модели (2.4) изображение получается просто интегрированием оригинала и функции рассеяния

точки аналоговыми средствами. Когда модель вводится в ЦВМ, изображение, оригинал и функция рассеяния точки дискретизируются на N^2, N^2 и N^4 отсчетов соответственно.

В непрерывно-дискретной модели (2.5) подразумевается, что оригинал непрерывен (поскольку он является объектом реального мира), а датчик, формирующий изображение, дискретен, т. е. продискретизирован на N^2 отсчетов. Таким образом, получается N^2 скалярных значений g_i и, следовательно, N^2 различных функций рассеяния точки $h_i(\xi, \eta)$.

Наконец, в дискретно-дискретной модели считается, что как оригинал f, так и изображение g представлены в виде N^2 одномерных векторов. Эти векторы могут быть образованы путем растрового сканирования двумерных функций либо с помощью какого-либо другого метода развертки, позволяющего получить все N^2 отсчетов. В данном случае четырехмерная функция $h(\xi, \eta, x, y)$ путем растрового сканирования преобразуется в двумерную матрицу, имеющую размер $N^2 \times N^2$.

В этой главе мы ограничимся обсуждением дискретно-дискретной модели, которую иногда будем видоизменять, вводя дополнительные упрощающие предположения. В частности, когда искажающее явление можно представить разделимой моделью, (2.6) может быть заменено выражением

$$[G] = [A][F][B], \tag{2.7}$$

где $[F]$ — дискретизированный вариант оригинала, $[G]$ — дискретизированный вариант изображения, а $[A]$ и $[B]$ — столбец и строка соответственно, выражающие искажение оригинала. Таким образом, все матрицы в (2.7) имеют размер $N \times N$, что существенно упрощает последующие вычисления.

2.2. Математическое представление

2.2.1. Представление изображения в ортогональном базисе

Различные аспекты цифровой обработки изображений нельзя рассматривать в отрыве от цифрового запоминания двумерных массивов чисел, представляющих отдельные значения яркости, полученные из исходной фотографии или сцены либо определенные с помощью телевизионной передающей трубки. Однако дикретизированное и квантованное изображение представляет собой просто матрицу положительных чисел (возможно, довольно большого размера), над которой в ЦВМ могут производиться линейные и нелинейные операции широкого класса. Поэтому цифровая обработка изображений может быть сведена к ана-

лизу манипуляций с большими матрицами, связанных с выполнением таких операций обработки, как преобразование, разложение, представление и реставрация. В данном разделе некоторые из таких операций исследуются с использованием аппарата матричной алгебры и, в частности, векторных внешних произведений.

Условимся рассматривать матрицу $[G]$ как дискретизированное и квантованное изображение; при этом i-я строка и j-й столбец матрицы соответствуют пространственным координатам x и y изображения $g(x, y)$. Таким образом, можно записать

$$[G] = SQ\{g(x, y)\}, \tag{2.8}$$

где нелинейный оператор $SQ\{\ldots\}$ выражает действие дискретизатора и квантователя. Примем, что $[G]$ имеет размер $N \times N$. Методы, пригодные для изображений, выражаемых неквадратными матрицами, рассматриваются в значительной части литературы; однако мы здесь воздержимся от их обсуждения, поскольку это не может существенно повлиять на излагаемые принципы. Обобщенное линейное разделимое преобразование матрицы изображения $[G]$ может быть записано в виде

$$[\alpha] = [U]^t [G] [V], \tag{2.9}$$

где $[\alpha]$ — изображение, подвергнутое унитарному преобразованию, $[U]$ и $[V]$ — унитарные операторы, а индекс t обозначает транспонирование матрицы. Унитарность $[U]$ и $[V]$ означает, что

$$[U][U]^t = [I], \tag{2.10а}$$

$$[V]^t [V] = [I]. \tag{2.10б}$$

Отсюда следует, что преобразование, обратное (2.9), можно записать в виде

$$[G] = [U][\alpha][V]^t. \tag{2.11}$$

Если раскрыть обозначения $[U]$ и $[V]$, записав

$$[U] = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \ldots\ \boldsymbol{u}_N], \tag{2.12}$$

$$[V] = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_N], \tag{2.13}$$

где $\boldsymbol{u}_i$ и $\boldsymbol{v}_i$ — векторы, образованные из столбцов $[U]$ и $[V]$, то получим

$$[G] = [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \ldots\ \boldsymbol{u}_N][\alpha]\begin{bmatrix} \boldsymbol{v}_1^t \\ \boldsymbol{v}_2^t \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{v}_N^t \end{bmatrix}. \tag{2.14}$$

Если матрицу α записать в виде суммы

$$[\alpha]=\begin{bmatrix}\alpha_{11} & 0 & 0 & \dots & 0\\ 0 & 0 & & & \\ \vdots & & & & \\ \vdots & & & & \\ 0 & & & & 0\end{bmatrix}+\begin{bmatrix}0 & \alpha_{12} & 0 & \dots & 0\\ 0 & & & & \\ \vdots & & & & \\ \vdots & & & & \\ 0 & \dots & 0 & & \end{bmatrix}+\dots, \qquad (2.15)$$

то отсюда следует, что

$$[G]=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}\boldsymbol{u}_i\boldsymbol{v}_j^t. \qquad (2.16)$$

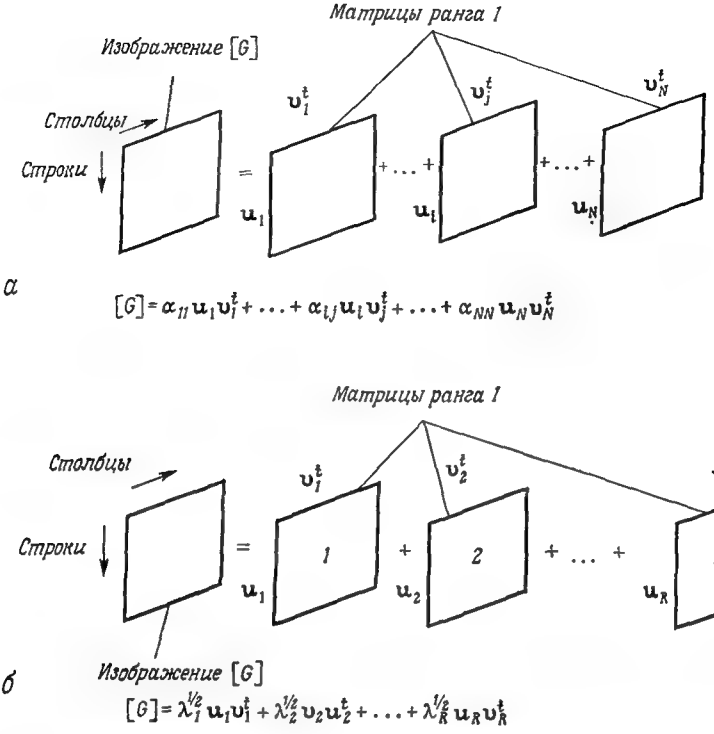
Внешнее произведение $\boldsymbol{u}_i\boldsymbol{v}_j^t$ можно интерпретировать как «изображение», так что сумма по всем комбинациям внешних произведений с соответствующими весами α_{ij} восстанавливает исходное изображение $[G]$. Для некоторых видов преобразований такой подход позволяет дать исключительно удобную интерпретацию аппроксимаций с сохранением выбранных коэффициентов.

В качестве графической иллюстрации этого процесса рассмотрим пример на фиг. 2.2, *а*. Здесь изображение $[G]$ разлагается в сумму N^2 матриц ранга 1 с соответствующими коэффициентами α_{ij}. Матрицы ранга 1 представляют двумерные базисные изображения, примерами которых служат тригонометрические функции для разложения по Фурье и бинарные функции для разложения по Уолшу.

Выбор преобразований, выражаемых матрицами $[U]$ и $[V]$, вполне произволен. Эти матрицы могут выбираться как из одних и тех же, так и из различных базисных функций. Так, например, матрица $[U]$ может быть составлена из функций Уолша, а $[V]$ — из функций Фурье, т. е. синусов и косинусов. В этом частном случае столбцы матрицы изображения разлагаются в ряд по функциям Уолша (прямоугольным функциям), а строки — в ряд по комплексным экспонентам (синусам и косинусам).

Хотя это и не следует явно из (2.16), при соблюдении определенных условий это выражение можно интерпретировать как разложение по сингулярным значениям (РСЗ), а в данном случае — как разложение изображения по его сингулярным векторам [1, 2]. Если $[\alpha]$ — диагональная матрица ранга R (т. е. если в $[\alpha]$ имеется лишь R положительных диагональных элементов), то

$$[G]=\sum_{i=1}^{R}\alpha_i\boldsymbol{u}_i\boldsymbol{v}_i^t. \qquad (2.17)$$



Фиг. 2.2. Разложение изображения по внешним произведениям: *а* — разложение по матрицам ранга 1; *б* — разложение по сингулярным значениям $[G]$.

Здесь α_i — квадратные корни из сингулярных значений $[G]\,[G]^t$ (т. е. $\alpha_i = \lambda_i^{1/2}$), а $\boldsymbol{u}_i$ и $\boldsymbol{v}_i$ — сингулярные векторы. Другой, более традиционный способ изложения строится на следующих определениях:

$$[G] = [U]\,[\Lambda]^{1/2}\,[V]^t \tag{2.18а}$$

и

$$[G]\,[G]^t = [U]\,[\Lambda]\,[V]^t, \tag{2.18б}$$

$$[G]^t\,[G] = [V]\,[\Lambda]\,[V]^t, \tag{2.18в}$$

где $[\Lambda]$ — диагональная матрица собственных значений $[G]\,[G]^t$, столбцы $[U]$ — собственные векторы $[G]\,[G]^t$, а столбцы $[V]$ — собственные векторы $[G]^t\,[G]$. Симметрия и квадратность $[G]\,[G]^t$ и $[G]^t\,[G]$ гарантируют действительность λ_i и ортогональность множеств собственных векторов $\{\boldsymbol{u}_i\}$ и $\{\boldsymbol{v}_i\}$.

Из (2.17) с очевидностью следует, что уменьшение R влечет за собой сокращение числа степеней свободы, определяющих изображение $[G]$. Располагая собственные значения в порядке монотонного убывания, мы получаем наиболее эффективное среднеквадратичное представление изображения с помощью наименьших (усеченных) множеств сохраняемых составляющих $\{\lambda_l^{1/2}, \boldsymbol{u}_l, \boldsymbol{v}_l\}$. Этот факт используется в следующем разделе. Кроме того, метод РСЗ обеспечивает удобный переход к псевдоинверсии матриц, что имеет важное значение для обсуждаемой ниже реставрации изображений в случае разделимой пространственно-зависимой функции рассеяния точки.

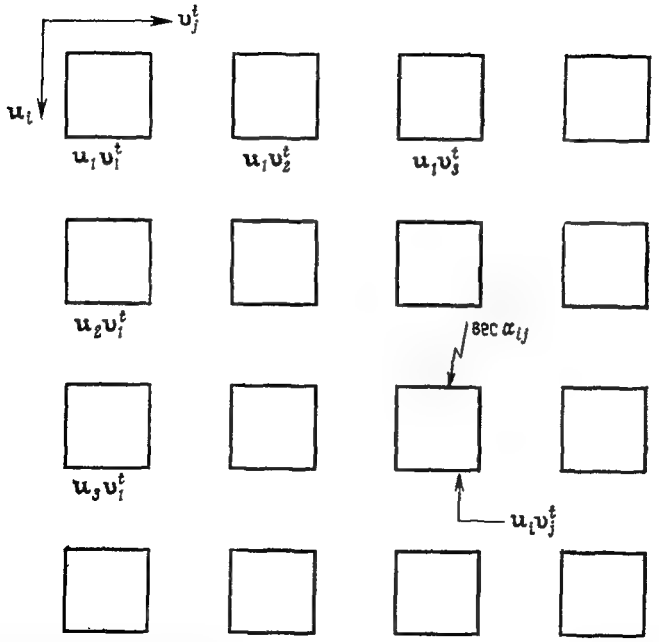
Резюмируя, заметим, что РСЗ состоит просто в определении сингулярных (или собственных) векторов для $[G][G]^t$ и $[G]^t[G]$. Затем полученные сингулярные векторы и значения могут быть использованы для эффективного (в среднеквадратичном смысле) представления изображения $[G]$. Тот, кто знаком с методом разложения на главные компоненты [1]), может отождествить матрицы $[G][G]^t$ и $[G]^t[G]$ с выборочными строчными и столбцовыми ковариационными матрицами; при этом РСЗ подобных статистически определенных процессов переходит в разложение изображения по собственным изображениям. Однако читателю следует воздержаться от вывода, что метод РСЗ совпадает с разложением Карунена — Лоэва. В действительности эти разложения не одинаковы, и их различия обсуждаются в последующих разделах.

Обращаясь вновь к фиг. 2.2, *б*, мы видим, что РСЗ изображения $[G]$ содержит лишь R членов с весовыми коэффициентами, которые выражаются корнями из соответствующих сингулярных значений. Входящие в это разложение матрицы ранга 1 также представляют собой ортогональные базисные изображения, но «настроенные» непосредственно на данное конкретное изображение, благодаря чему они обеспечивают единственное разложение по оптимальному базису.

2.2.2. Примеры представления изображений

Для целей сокращения полосы частот или кодирования при цифровой обработке изображений мы ищем эффективные средства накопления, минимизирующие ошибку усечения при значительном сокращении полосы частот. Методы кодирования преобразований основаны на переходе в преобразованную область с помощью матрицы $[\alpha]$, определяемой выражением (2.9), и усечении, выполняемом путем придания определенным α_{ij} нуле-

[1]) См. [5], раздел II, и [3], гл. 6. — *Прим. перев.*



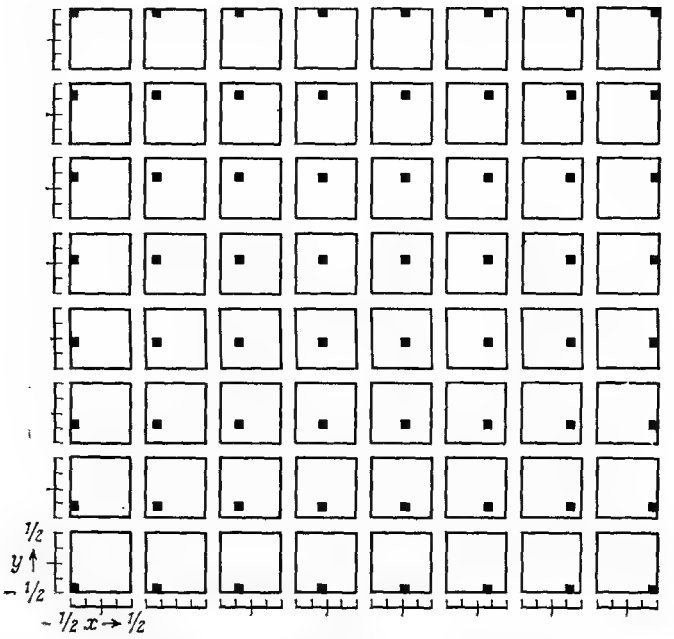
Фиг. 2.3. Формат ортогонального разложения общего вида по внешним произведениям с разделимым ядром.

вых значений. Таким образом, в (2.16) будет меньше чем N^2 членов, и это выражение принимает вид

$$[G_K] = \sum_{i,\ j \in \{K\}} \sum \alpha_{ij} \boldsymbol{u}_i \boldsymbol{v}_j^t, \tag{2.19}$$

где $\{K\}$ — некоторое подмножество коэффициентов из общего числа N^2.

Примеры ортогональных разложений общего вида с разделимым ядром удобнее всего представлять в форме, показанной на фиг. 2.3. Изображения внешних произведений представлены здесь в виде матрицы, которая служит для иллюстрации того, что изображение $[G]$ может быть разложено на показанные квадраты $\boldsymbol{u}_i\boldsymbol{v}_j^t$ с весами α_{ij}. Для тех ортогональных систем базисных векторов, которые содержат чисто единичный вектор 1, множество действительных одномерных базисных векторов располагается в первом столбце и первой строке таблицы, представленной на фиг. 2.3. Множество возможных ортогональных систем для разложения изображения бесконечно; здесь обсуждаются только простейшие из них. На фиг. 2.4—2.7 представ-



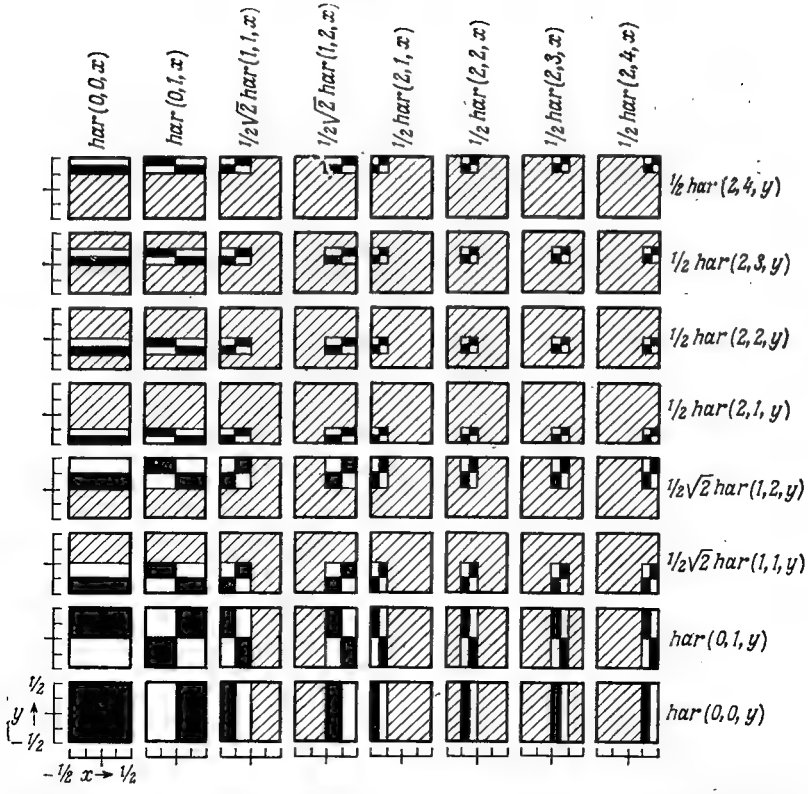
Фиг. 2.4. Разложение по внешним произведениям для преобразования тождественности (черным и белым показаны значения +1 и 0 соответственно).

лены изображения внешних произведений, которые иллюстрируют преобразование тождественности, преобразования Хаара, Адамара — Уолша и еще один вариант преобразования Адамара [3]. Во всех этих случаях $[U]=[V]$, т. е. преобразования по строкам и столбцам одинаковы. На фиг. 2.4—2.7 матрица $[U]$ принимает следующие значения [4]:

а) преобразование тождественности

$$[U]=\begin{bmatrix} 1&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&1 \end{bmatrix}$$

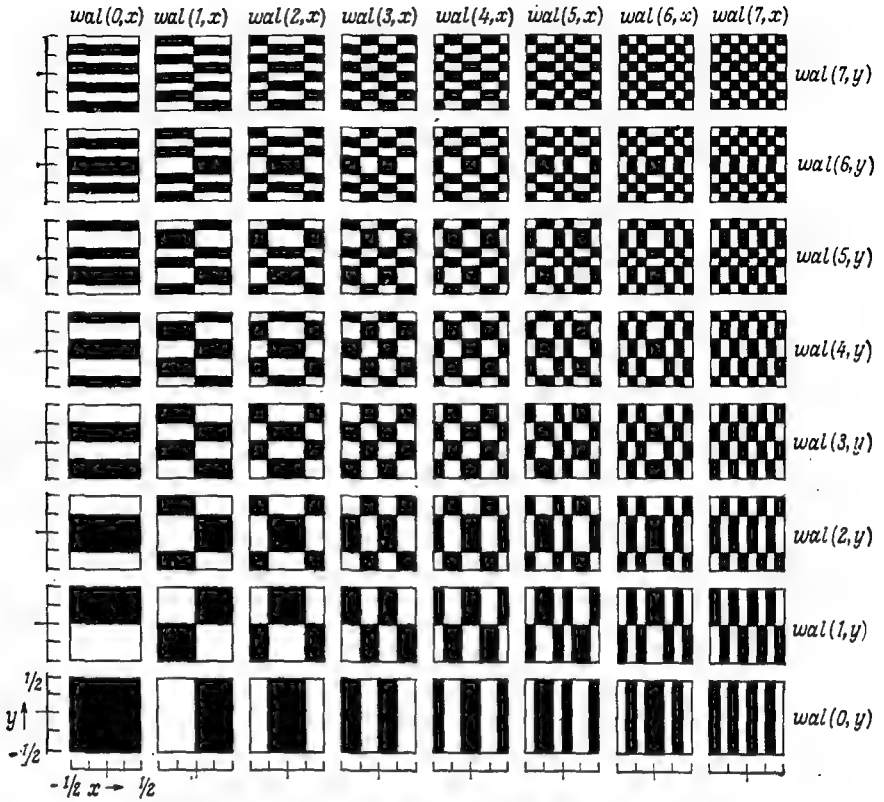
(см. фиг. 2.4),



Фиг. 2.5. Разложение по внешним произведениям для преобразования Хаара (черным и белым показаны значения $+1$ и -1 соответственно, заштрихованы значения 0).

б) преобразование Хаара

$$[U]=\begin{bmatrix} 1 & -1 & 2 & 0 & 2 & 0 & 0 & 0 \\ 1 & -1 & 2 & 0 & -2 & 0 & 0 & 0 \\ 1 & -1 & -\sqrt{2} & 0 & 0 & 2 & 0 & 0 \\ 1 & -1 & -\sqrt{2} & 0 & 0 & -2 & 0 & 0 \\ 1 & 1 & 0 & 2 & 0 & 0 & 2 & 0 \\ 1 & 1 & 0 & 2 & 0 & 0 & -2 & 0 \\ 1 & 1 & 0 & -\sqrt{2} & 0 & 0 & 0 & 2 \\ 1 & 1 & 0 & -\sqrt{2} & 0 & 0 & 0 & -2 \end{bmatrix}$$



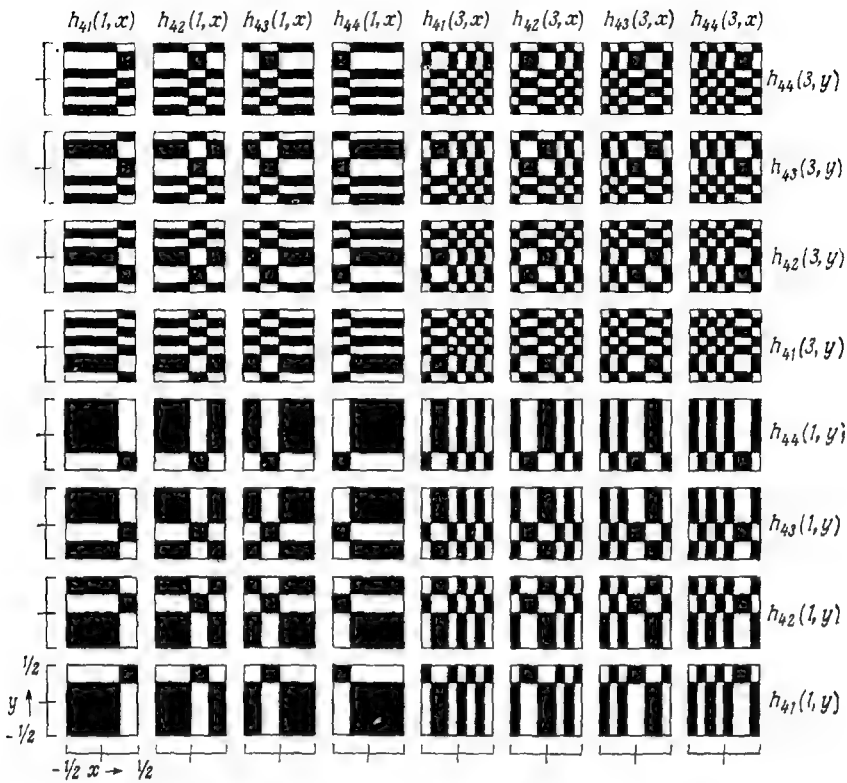
Фиг. 2.6. Разложение по внешним произведениям для преобразования Адамара — Уолша (черным и белым показаны значения $+1$ и -1 соответственно).

или $[U]=[\mathcal{H}]$ (см. фиг. 2.5, где ради упрощения полутонового представления опущен нормирующий коэффициент $\sqrt{2}$),

в) преобразование Адамара — Уолша

$$[U]=\begin{bmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix}$$

(см. фиг. 2.6),

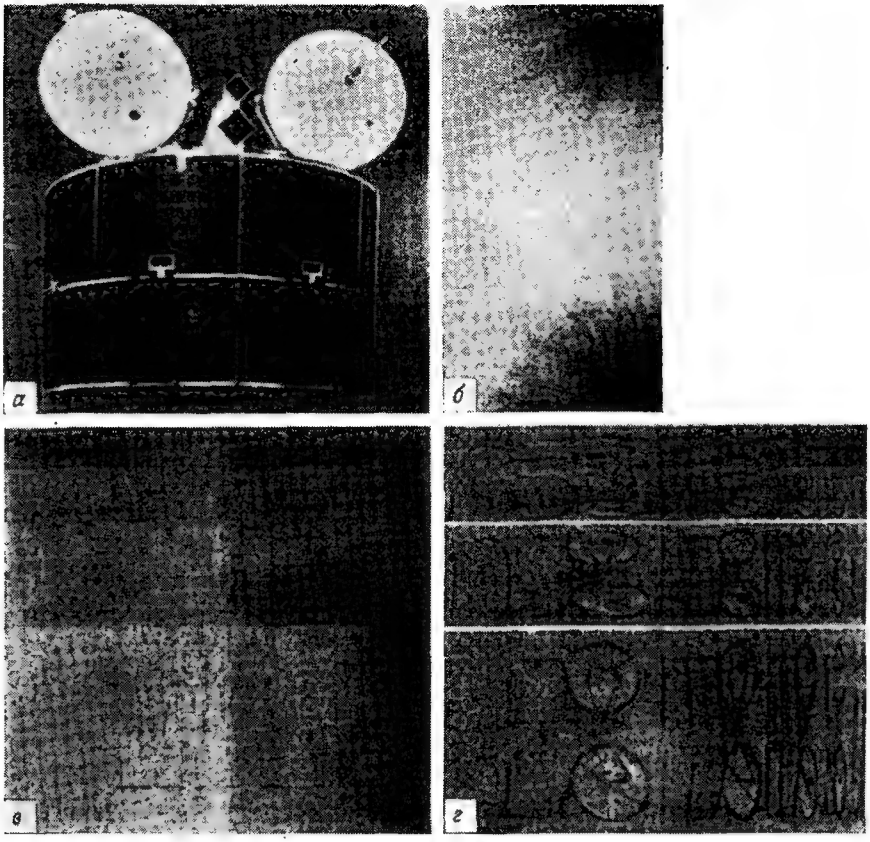


Фиг. 2.7. Разложение по внешним произведениям для варианта преобразования Адамара (черным и белым показаны значения +1 и —1 соответственно).

г) вариант преобразования Адамара

$$[U]=\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

(см. фиг. 2.7),

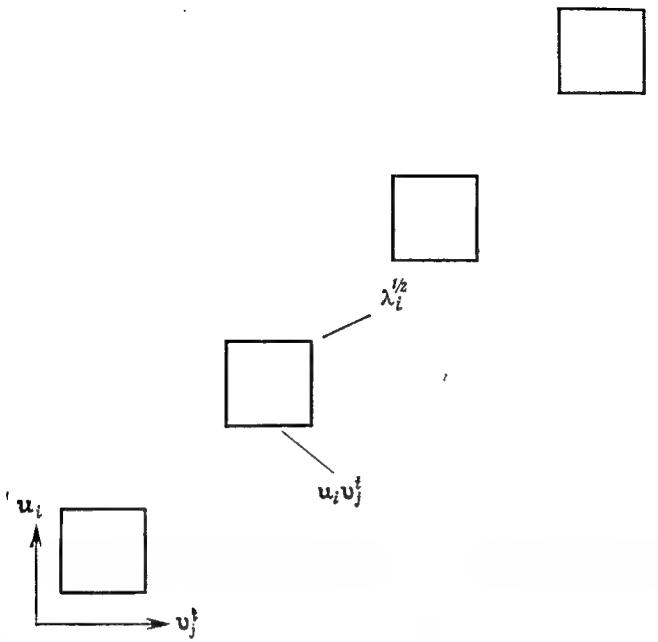


Фиг. 2.8. Разложение по внешним произведениям изображения спутника: *а* — исходное изображение; *б* — преобразование Фурье; *в* — преобразование Адамара; *г* — преобразование Хаара.

д) преобразование Фурье

$$[U] = \overset{u\downarrow}{\underset{}{}}\overset{x\rightarrow}{\underset{}{}} e^{\frac{2\pi i u x}{N}} = \exp\left\{\frac{2\pi i}{8}\begin{bmatrix} 0&0&0&0&0&0&0&0\\ 0&1&2&3&4&5&6&7\\ 0&2&4&6&0&2&4&6\\ 0&3&6&1&4&7&2&5\\ 0&4&0&4&0&4&0&4\\ 0&5&2&7&4&1&6&3\\ 0&6&4&2&0&6&4&2\\ 0&7&6&5&4&3&2&1 \end{bmatrix}\right\}$$

или $[U] = [\mathscr{F}]$ (на фигуре не показано.)

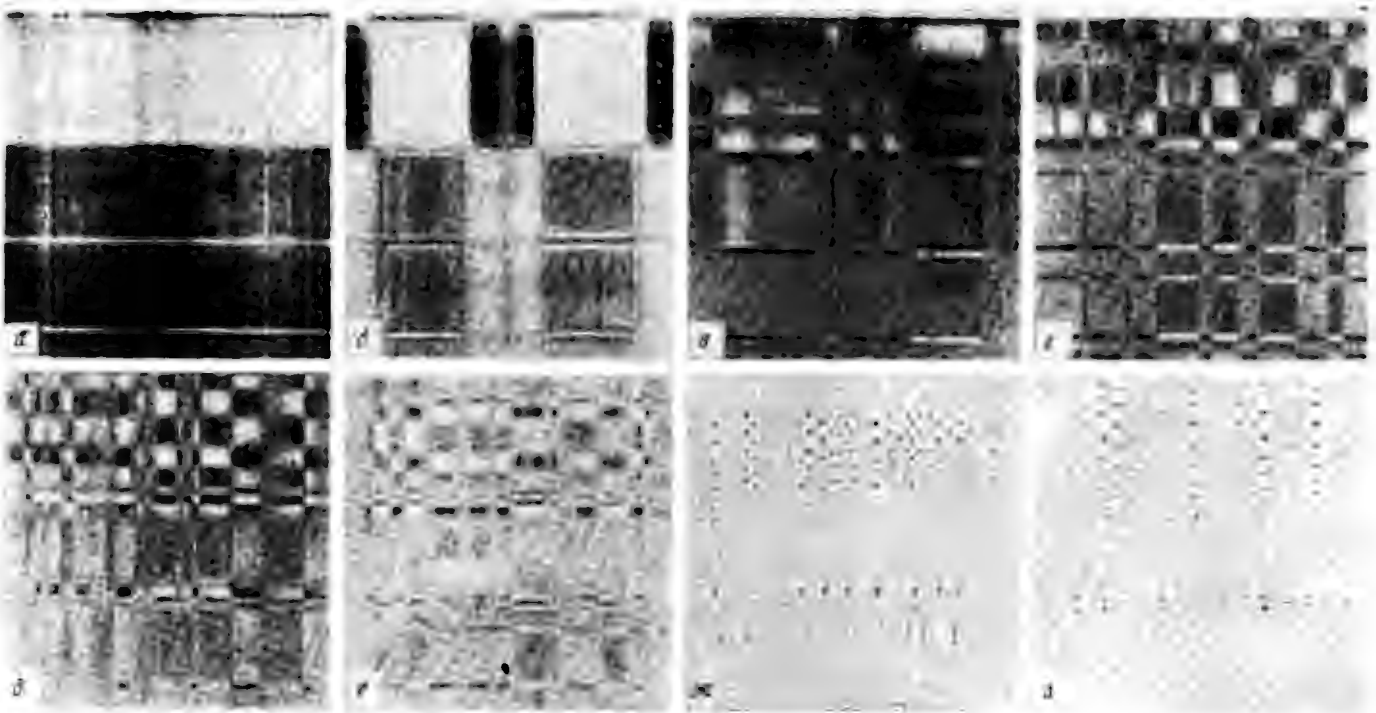


Фиг. 2.9. Формат разложения общего вида по сингулярным значениям.

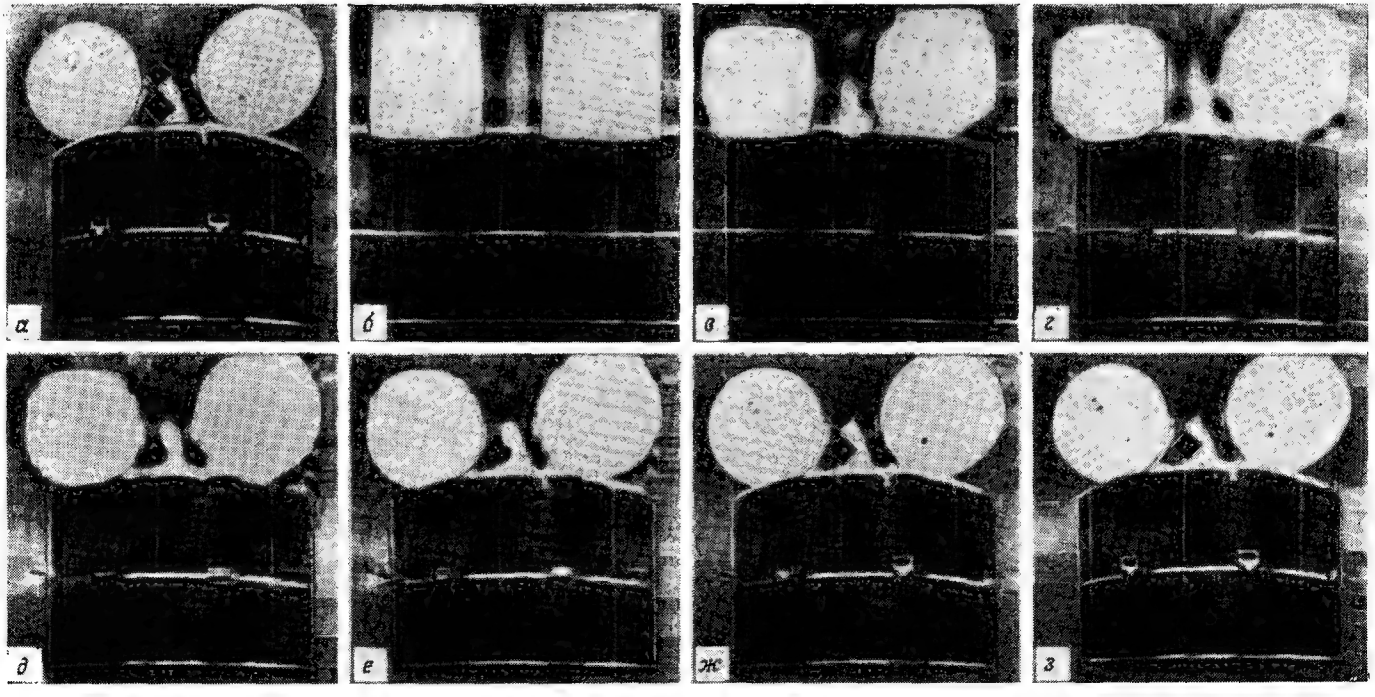
Разумеется, возможны различные сочетания ортогональных систем, при которых $[U] \neq [V]$, однако из-за недостатка места они здесь не рассматриваются.

Чтобы проиллюстрировать применение подобных разложений, рассмотрим точечные решетки более высокой размерности, построенные на основе реальных изображений. На фиг. 2.8 представлены матрицы $[\alpha]$, которые для данного конкретного изображения соответствуют разложениям по внешним произведениям в случае преобразования тождественности и преобразований Фурье, Уолша и Хаара. Изображения в преобразованных областях, соответствующие матрицам $[\alpha]$, представлены в виде логарифмов амплитуд их элементов, что позволило согласовать воспроизводимые величины с узким динамическим диапазоном фотопленки. Таким образом, читатель фактически видит матрицу $[\alpha']$, в которой $\alpha'_{ij} = \log(|\alpha_{ij}| + 1)$.

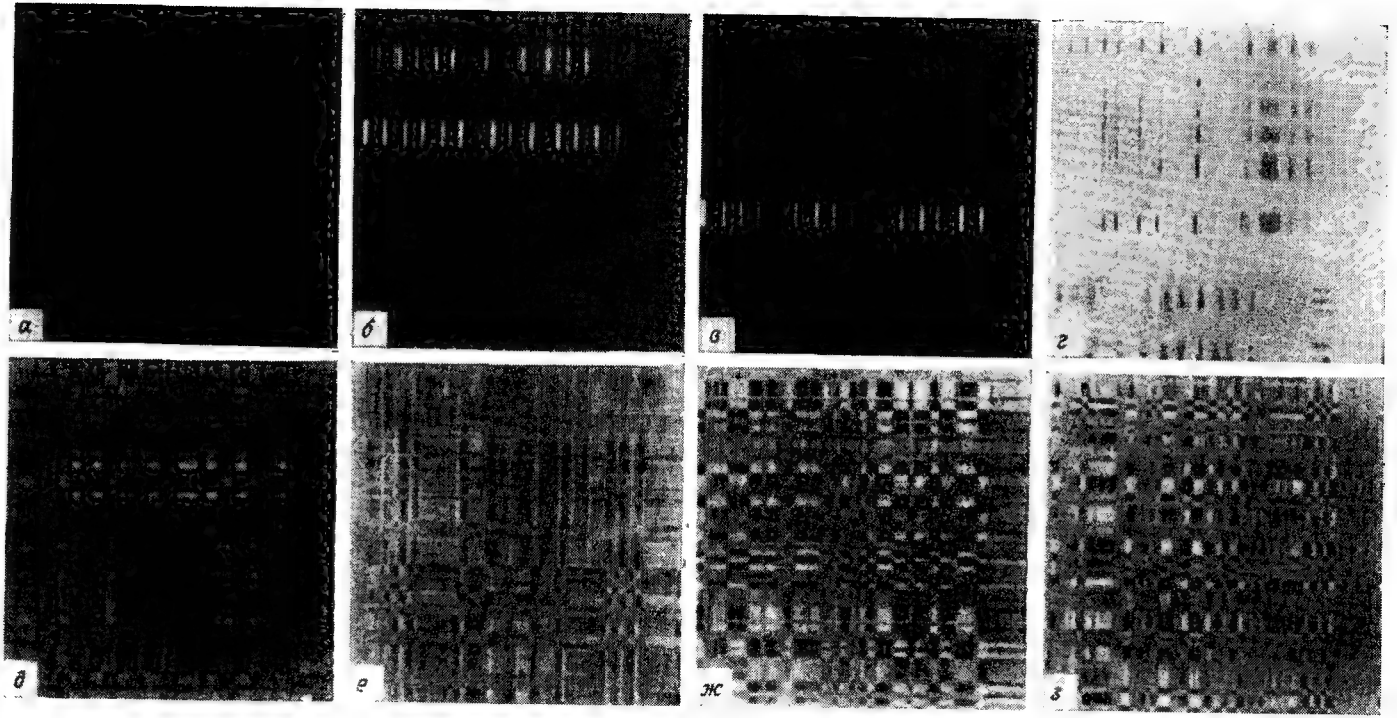
Как видно из фиг. 2.8, энергия изображения стремится сосредоточиться на нескольких избранных коэффициентах α_{ij}; следовательно, чтобы обеспечить хорошее воспроизведение исходной картины, требуется сохранить лишь немногие члены разложения. Поскольку обсуждаемые разложения унитарны, энергия



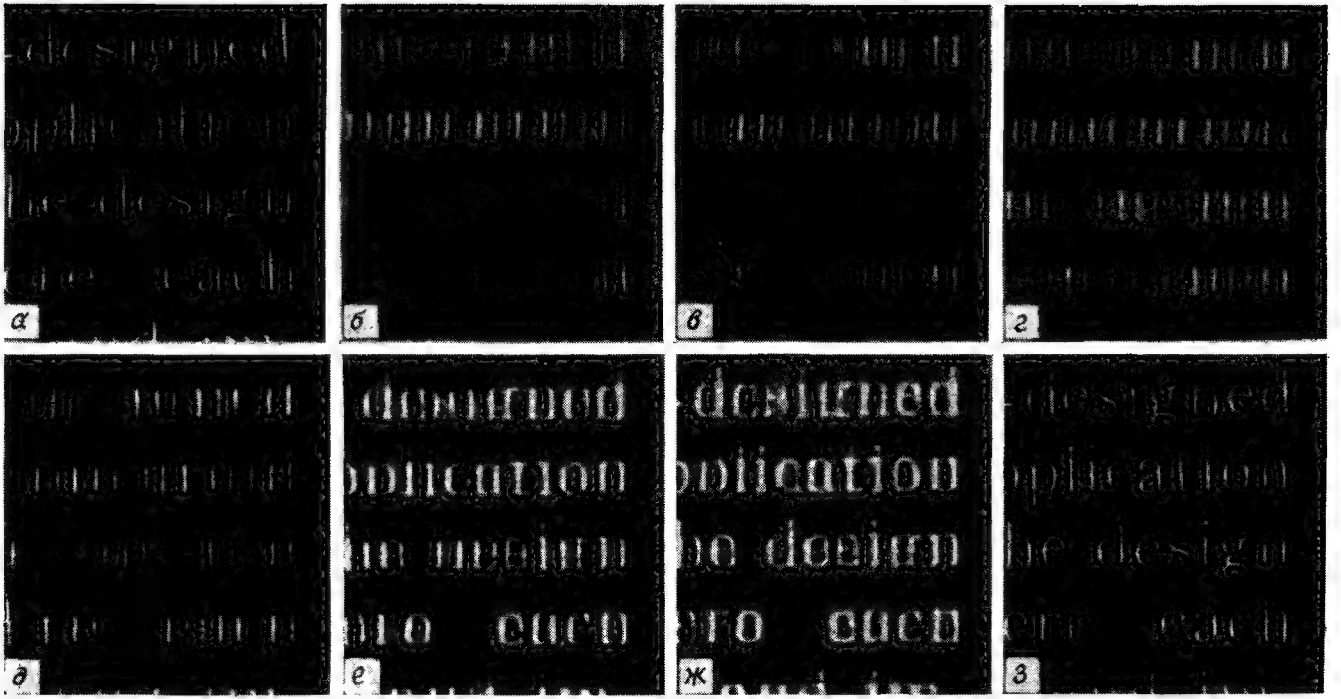
Фиг. 2.10. Выбранные собственные значения внешних произведений сингулярных векторов: $а - \boldsymbol{u}_1 \boldsymbol{v}_1^t$; $б - \boldsymbol{u}_2 \boldsymbol{v}_2^t$; $в - \boldsymbol{u}_4 \boldsymbol{v}_4^t$; $г - \boldsymbol{u}_6 \boldsymbol{v}_6^t$; $д - \boldsymbol{u}_8 \boldsymbol{v}_8^t$; $е - \boldsymbol{u}_{10} \boldsymbol{v}_{10}^t$; $ж - \boldsymbol{u}_{20} \boldsymbol{v}_{20}^t$; $з - \boldsymbol{u}_{30} \boldsymbol{v}_{30}^t$.



Фиг. 2.11. Выбранные частные суммы РСЗ изображения спутника: *а* — исходное изображение G_{128}; *б* — G_2; *в* — G_4; *г* — G_6; *д* — G_8; *е* — G_{10}; *ж* — G_{20}; *з* — G_{30}.



Фиг. 2.12. Выбранные собственные изображения внешних произведений сингулярных значений текста: $а — \boldsymbol{u}_1\boldsymbol{v}_1^t$; $б — \boldsymbol{u}_2\boldsymbol{v}_2^t$; $в — \boldsymbol{u}_4\boldsymbol{v}_4^t$; $г — \boldsymbol{u}_6\boldsymbol{v}_6^t$; $д — \boldsymbol{u}_8\boldsymbol{v}_8^t$; $е — \boldsymbol{u}_{10}\boldsymbol{v}_{10}^t$; $ж — \boldsymbol{u}_{20}\boldsymbol{v}_{20}^t$; $з — \boldsymbol{u}_{30}\boldsymbol{v}_{30}^t$.



Фиг. 2.13. Выбранные частные суммы РСЗ текста: *а* — исходное изображение G_{128}; *б* — G_2; *в* — G_4; *г* — G_6; *д* — G_8; *е* — G_{10}; *ж* — G_{20}; *з* — G_{30}.

изображения преобразуется так, что несколько больших коэффициентов α_{ij} представляют большую долю энергии изображения. Выбор подходящего разложения по внешним произведениям и передача выбранных коэффициентов составили содержание исследований по цифровому кодированию изображений посредством преобразований, при которых удалось уменьшить требуемую емкость памяти приблизительно в 10 раз [5].

В случае разложения по сингулярным значениям недиагональные коэффициенты α_{ij}, у которых $i \neq j$, равны нулю. Фиг. 2.9 иллюстрирует ортогональную систему функций, используемых при РСЗ. Рассматривая РСЗ как средство представления, сохраняющее лишь K таких внешних произведений, мы находим, что

$$[G_K] = \sum_{i=1}^{K} \lambda_i^{1/2} \boldsymbol{u}_i \boldsymbol{v}_i^t, \tag{2.20}$$

причем норма разности $[G]$ и $[G_K]$ составляет

$$\|[G] - [G_K]\| = \sum_{i=K+1}^{R} \lambda_i. \tag{2.21}$$

Следовательно, монотонно убывающий порядок собственных значений минимизирует норму, или среднеквадратичную ошибку усечения.

На фиг. 2.10—2.13 показаны различные РСЗ изображений. Чтобы проиллюстрировать разложение изображений на их базисные двумерные составляющие, здесь представлены (в виде амплитуд их составляющих) отдельные выбранные изображения внешних произведений сингулярных векторов. Наряду с отдельными сингулярными векторными матрицами показаны также различные частные суммы этих матриц, чтобы продемонстрировать совокупное влияние отдельных внешних произведений на восстановление полного изображения. Очевидно, что для сохранения в изображении важной информации требуется лишь небольшое количество членов разложения.

2.2.3. Разделимость и пакетные операторы

До сих пор при обсуждении использовалось неявное предположение о прямоугольной разделимости. Повторяемое здесь выражение (2.11)

$$[G] = [U][\alpha][V]^t$$

означает, что строки матрицы $[\alpha]$ обрабатываются посредством $[V]^t$, а столбцы $[\alpha]$ — посредством $[U]$. Поскольку $[U]$ и $[V]$ могут выбираться произвольно, такое предположение о разделимости приводит к некоторым последствиям, которые

рассматриваются ниже. В частности, это предположение влечет за собой ограничение общности линейной взаимосвязи между преобразованной областью $[\alpha]$ и изображением $[G]$. Наиболее общая линейная система, связывающая $[\alpha]$ и $[G]$, определяется пакетным (stacking) оператором [6], который отображает матрицы $[\alpha]$ и $[G]$ в $(N^2 \times 1)$ векторы $\boldsymbol{\alpha}$ и $\boldsymbol{g}$ соответственно. При этом получаем

$$\boldsymbol{g} = [H]\boldsymbol{\alpha}. \qquad (2.22)$$

В этом случае $[H]$ имеет размер $N^2 \times N^2$, т. е. обладает N^4 степенями свободы, тогда как в случае разделимости в соответствии с (2.11) допускается лишь $2N^2$ степеней свободы. Однако если матрица $[H]$ используется для представления операций при условии разделимости, то она переходит в кронекерово, или прямое, произведение матриц вида [7]

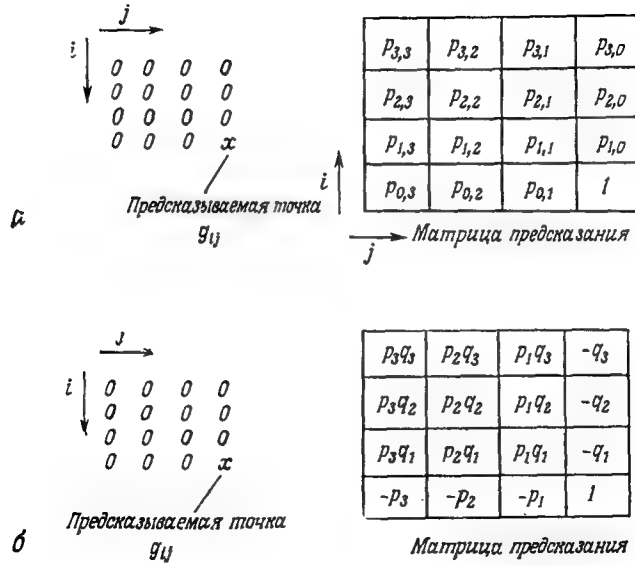
$$[H] = [U] \otimes [V] \qquad (2.23)$$

и приобретает блочную структуру. Мы увидим, что предположение о разделимости позволяет получить существенные вычислительные преимущества, однако оно сводится к ряду априорных допущений, выполнение которых не может быть обеспечено на интуитивной основе. Следует подчеркнуть, что пакетный оператор позволяет записать более общую линейную взаимосвязь между областями $[\alpha]$ и $[G]$, чем та, которая возможна в случае разделимости.

2.3. Кодирование изображений

2.3.1. Пространственное кодирование изображений

Цель кодирования изображений состоит в эффективном сокращении количества битов, необходимого для представления изображения, и сохранении при этом некоторого уровня верности воспроизведения кодированного изображения по отношению к исходному. Существуют два общих подхода, один из которых охватывает методы кодирования изображения в исходной (пространственной) области, а другой — в преобразованной области. Однако многие из пространственных методов можно представить как методы кодирования двумерных преобразований изображения, и именно такой подход используется в этой главе. На фиг. 2.14 представлены отсчеты некоторого изображения и поясняются общие принципы дифференциальной ИКМ (ДИКМ),



Фиг. 2.14. Дифференциальная ИКМ: *а* — неразделимая дифференциальная ИКМ; *б* — разделимая дифференциальная ИКМ.

Для $(n \times n)$-предсказателя получаем в положении (i, j) разность Δ_{ij} в виде

$$\Delta_{i,j} = g_{i,j} - \sum_{k=1}^{n} p_{k,0} g_{i-k,j} - p_{0,l} g_{i,j-l} - \sum_{k=1}^{n} \sum_{l=1}^{n} p_{k,l} g_{i-k,j-l}, \quad (2.24)$$

которая затем кодируется и передается. При неразделимой дифференциальной ИКМ (фиг. 2.14, *а*) имеется $n^2 + 2n$ коэффициентов предсказателя $p_{k,l}$ и соответственно $n^2 + 2n$ степеней свободы предсказателя. Часто такой предсказатель называют предсказателем $(n^2 + 2n)$-го порядка. В случае разделимой дифференциальной ИКМ (фиг. 2.14, *б*) предсказание отсчета $g_{i,j}$ дает разность

$$\Delta_{i,j} = g_{i,j} - \sum_{k=1}^{n} p_k g_{i-k,j} - \sum_{l=1}^{n} q_l g_{i,j-l} + \sum_{k=1}^{n} \sum_{l=1}^{n} p_k q_l g_{i-k,j-l}, \quad (2.25)$$

и мы видим, что здесь используются лишь $2n$ однозначных коэффициентов предсказания, т. е. имеются $2n$ степеней свободы (предсказатель $2n$-го порядка). В случае когда изображение

имеет разделимую статистику [5], предсказатель описывается двумя треугольными матрицами:

$$[Q]=\begin{bmatrix}
1 & & & & & & & & & & & & & & \\
-q_1 & 1 & & & & & & & & & & & & & \\
-q_2 & -q_1 & 1 & & & & & & & & & & & & \\
\cdot & & & 1 & & & & & & & & & & & \\
\cdot & & & & 1 & \cdot & & & & & & & & & \\
\cdot & & & & & 1 & & & & & & & & & \\
-q_n & \cdot & \cdot & \cdot & -q_2 & -q_1 & 1 & & & & & & & & \\
 & \cdot & & & & & & \cdot & & & & & & & \\
 & & \cdot & & & & & & \cdot & & & & & & \\
 & & & & \cdot & \cdot & \cdot & & & \cdot & & & & & \\
 & & & -q_n & \cdot & \cdot & \cdot & -q_2 & & -q_1 & 1 & & & & \\
 & & & & -q_n & \cdot & \cdot & \cdot & \cdot & -q_2 & -q_1 & 1 & & &
\end{bmatrix} \tag{2.26}$$

для предсказания по столбцам и

$$[P]^t=\begin{bmatrix}
1 & -p_1 & -p_2 & \cdot & \cdot & \cdot & -p_n & & & & & & & \\
 & 1 & -p_1 & & & & & \cdot & & & & & & \\
 & & 1 & & & & & & \cdot & & & & & \\
 & & & 1 & & & & & & \cdot & & & & \\
 & & & & 1 & & & & & & -p_n & & & \\
 & & & & & 1 & -p_1 & -p_2 & \cdot & \cdot & \cdot & -p_n & & \\
 & & & & & & \cdot & & & & \cdot & \cdot & & \\
 & & & & & & & \cdot & & & \cdot & \cdot & & \\
 & & & & & & & & \cdot & & -p_2 & \cdot & & \\
 & & & & & & & & & 1 & -p_1 & -p_2 & & \\
 & & & & & & & & & & 1 & -p_1 & & \\
 & & & & & & & & & & & 1 & &
\end{bmatrix} \tag{2.27}$$

для предсказания по строкам [8, 9]. В этом случае разностное изображение $[\Delta]$ принимает вид

$$[\Delta]=[Q][G][P]^t, \tag{2.28}$$

и именно оно кодируется и передается по каналу. (Ради упрощения обсуждения мы игнорировали здесь граничные условия.) Разделимый предсказатель для ДИКМ дает адекватный результат, когда изображения оказываются статистически разделимыми и описываются горизонтальной и вертикальной ковариационными матрицами. Модели таких систем могут включать

в себя марковские процессы n-го порядка в двух измерениях. Изображение $[\Delta]$ принимает форму матрицы некоррелированных отсчетов, которая может быть закодирована более эффективно, чем матрица $[G]$ исходных коррелированных отсчетов (элементов изображения).

Из приведенных рассуждений следует, что с помощью раздельных по столбцам и строкам матричных операций изображение $[G]$, описываемое процессами определенного типа, может быть преобразовано в новую матрицу, более удобную для кодирования с целью сокращения полосы частот. В следующем разделе столбцовые и строчные матричные операции используются для перехода в область, обеспечивающую более эффективное кодирование изображения. Однако соответствующие операторы выражаются при этом матрицами унитарных преобразований, а не треугольными матрицами.

2.3.2. Кодирование изображений посредством преобразований

Кодирование изображений посредством преобразований достаточно подробно обсуждалось в литературе; хорошие обзоры по этому вопросу были представлены Уинтцем [5] и Хуангом и др. [10]. Главная цель такого кодирования состоит в переводе изображения в некоторую обратимую форму, более удобную для применения определенных методов кодирования. Обычно при этом получают данные с меньшей корреляцией, чем у элементов исходного изображения, и процесс сокращения полосы частот становится реализуемым с помощью кодеров без памяти, действующих в преобразованной области. При анализе изображений часто исходят из статистических характеристик второго порядка и предположения о стационарности, а для описания соответствующих процессов пользуются ковариационными матрицами.

В этом разделе мы рассмотрим вкратце ряд методов кодирования в преобразованной области и укажем некоторые характеристики каждого из них. Сводка наиболее часто встречающихся преобразований представлена в табл. 2.1, которая служит основой дальнейшего изложения. Во всех рассмотренных случаях область (матрица данных), используемая для кодирования и передачи, будет описываться матрицей $[\alpha]$, т. е. двумерным преобразованием матрицы изображения $[G]$. Необходимо заметить, что приведенную таблицу не следует считать исчерпывающей, и, очевидно, различные сочетания перечисленных преобразований позволяют строить всевозможные методы кодирования посредством преобразований, исходя из допустимых структур памяти, средств обработки и синхронизации и относящихся к

Таблица 2.1

Преобразования, применяемые при кодировании

	Название преобразования	Матричное представление	Неизвестные параметры (вычисляемые и передаваемые)	Алгоритмическая реализация (порядок числа операций)	Литература
1	Разложение по сингулярным значениям	$[G][G]^t = [U][\Lambda][U]^t$ $[G]^t[G] = [V][\Lambda][V]^t$ $[\alpha] = [\Lambda^{1/2}] = [U]^t[G][V]$	Базисные векторы	N^3 N^3 $2N^3$ } вычисляются для каждого изображения	[1, 2]
2	Карунена — Лоэва, Хотеллинга, по главным компонентам, факторный анализ	$[\Phi_x] = [E_x][\Lambda_x][E_x]^t$ $[\Phi_y] = [E_y][\Lambda_y][E_y]^t$ $[\alpha] = [E_x]^t[G][E_y]$	Коэффициенты преобразования	N^3 N^3 } вычисляются один раз $2N^3$	[5, 10]
3	Косинусное	$[\alpha] = [\cos]^t[G][\cos]$	То же	$2N^2\log_2 N$ [1,2)]	[11]
4	Фурье	$[\alpha] = [\mathscr{F}][G][\mathscr{F}]$	» »	$2N^2\log_2 N$ (комплексных) [1,2)]	[3]
5	Наклонное	$[\alpha] = [S]^t[G][S]$	» »	$2N^2\log_2 N$ [2)]	[12]
6	В дискретном линейном базисе	$[\alpha] = [\mathrm{DLB}]^t[G][\mathrm{DLB}]$	» »	$2N^2\log_2 N$ (целочисленных) [1,2)]	[13]
7	Уолша	$[\alpha] = [W][G][W]$	» »	$2N^2\log_2 N$ (сложения) [2)]	[14]
8	Хаара	$[\alpha] = [\mathscr{H}]^t[G][\mathscr{H}]$	» »	$2(N-1)$ [2)]	[3]
9	Гибридное	$[\alpha] = [Q][G][W]$	Предсказанные значения одномерного преобразования	$N^2\log_2 N + nN^2$ [3)]	[15]
10	ДИКМ	$[\alpha] = [Q][G][P]^t$	Предсказанные значения	$2nN^2$ [3)]	[9]

[1)] Когда $N \neq 2^k$, объем вычислений возрастает.

[2)] $N = 2^k$.

[3)] n пропорционально порядку предсказателя

ним ограничений. Кроме того, напрашивается интуитивный вывод, что эффективное устранение корреляции убывает от верхних строк таблицы к нижним. Хотя такой вывод может оказаться полезным, при подобных обобщениях следует всегда проявлять осторожность, поскольку, например, в данном случае два последних преобразования (гибридное и ДИКМ) позволяют обеспечить такую же декорреляцию, что и другие методы преобразования, при увеличении порядка предсказателя n.

Рассмотрим отдельные строки табл. 2.1. Метод разложения по сингулярным значениям пригоден для применения только при условии, что на передающей стороне (у источника изображения) имеется возможность выполнить весьма большой объем вычислений, поскольку необходимо вычислить сингулярные векторы, которые однозначно определяются обрабатываемым изображением, и эти вычисления требуют в целом выполнения $4N^3$ операций. Полученные сингулярные (или собственные) векторы, в которые в качестве постоянных входят собственные значения, пригодны теперь для кодирования и передачи. Вследствие обширных вычислений, необходимых для выполнения РСЗ, подобные методы практически пригодны только для запоминания изображений в мощных вычислительных системах. Усеченная форма $[G_K]$, выражаемая (2.19), является оптимальной с точки зрения среднеквадратичного приближения и обеспечивается только усеченным разложением по сингулярным значениям. Следовательно, если исходить из записи слов в машинную память с плавающей запятой (сохраняющей точность в противоположность записи в обычной кодовой форме), РСЗ оказывается единственным преобразованием, оптимальным в смысле среднеквадратичного критерия верности. При заданном K минимизируется следующее выражение:

$$\text{наименьшая квадратичная ошибка} = \|[G] - [G_K]\|. \qquad (2.29)$$

Другой подход к кодированию изображений, учитывающий их статистику и минимизирующий среднеквадратичную ошибку для стационарных изображений, представлен классом преобразований, описанных Каруненом — Лоэвом и Хотеллингом и называемых разложением по главным компонентам или факторным анализом. Этот класс преобразований представлен во второй строке таблицы, и, как видно из приведенных формул, здесь используется дополнительное предположение о разделимости стационарных статистических характеристик второго порядка. Выражение для минимизируемой среднеквадратичной ошибки записывается в виде

$$\text{среднеквадратичная ошибка} = \mathscr{E}\{\|[G] - [G_K]\|^2\}, \qquad (2.30)$$

где $\mathscr{E}\{\ldots\}$ — двумерный разделимый оператор математического ожидания, который усредняет рассматриваемые разделимые статистические характеристики, описывающие данный тип изображения. Из предположения о разделимости вытекает требование, чтобы статистические характеристики по горизонтали и вертикали были разделимыми (которое отнюдь не обязательно удовлетворяется). Ковариационные матрицы, описывающие столбцовые (вертикальные) и строчные (горизонтальные) статистические характеристики, выражаются соответственно как $[\Phi_x]$ и $[\Phi_y]$. Если для рассматриваемого класса изображений справедлива стационарная разделимая модель, то имеет смысл вычислять собственные векторы ковариационной матрицы (что выполняется однократно) с помощью большой вычислительной системы, удаленной от источника изображения и передатчика. Таким образом,

$$[\Phi_x] = [E_x][\Lambda_x][E_x]^t, \tag{2.31а}$$

$$[\Phi_y] = [E_y][\Lambda_y][E_y]^t, \tag{2.31б}$$

причем для определения каждой из ортогональных матриц $[E_x]$ и $[E_y]$ требуется порядка N^3 вычислений. (Здесь $[\Lambda_x]$ и $[\Lambda_y]$ — диагональные матрицы, диагональные элементы которых представляют собой собственные значения ковариационных матриц.) В этом случае появляется возможность реализовать (либо ускорить) аппаратурное разложение изображения $[G]$ на коэффициенты разложения, определяемые как

$$[\alpha] = [E_x]^t [G][E_y] \tag{2.32}$$

у источника изображения или передатчика. Для этой процедуры требуется выполнение $2N^3$ арифметических операций, что составляет экономию в два раза по сравнению с методом РСЗ в расчете на одно изображение. Важно отметить, что двумерное преобразование Карунена — Лоэва, определяемое (2.32), дает в результате полную матрицу $[\alpha]$, элементы которой таковы, что *в среднем* наибольшая часть энергии изображения сосредоточена в наименьшем количестве коэффициентов α_{ij}. Разумеется, действительные характеристики алгоритма такого типа будут хорошими с точки зрения психофизики явлений зрительного восприятия лишь в той степени, в какой статистическая модель будет соответствовать передаваемому детерминированному изображению.

Модели, используемые для описания разделимых стационарных процессов, встречающихся при генерировании изображений, часто определяются посредством марковских цепей; в этом случае ковариационные матрицы $[\Phi_x]$ и $[\Phi_y]$ являются тёплицевыми матрицами и имеют элементы $\Phi_{x\,ij} = \rho_x^{|i-j|}$ и $\Phi_{y\,ij} = \rho_y^{|i-j|}$.

Поскольку такие матрицы преобразуются в диагональные с помощью синусов или косинусов соответствующей частоты и фазы [16], для дискретной аппроксимации разложения этого типа больше всего подходит косинусное преобразование, занимающее третью строку табл. 2.1. Дополнительное преимущество этого преобразования состоит в наличии быстрого алгоритма реализации ($2N^2 \log_2 N$ операций вместо $2N^3$), а также в том, что здесь не требуется разложение по собственным значениям. Таким образом, это преобразование является детерминированным и может служить основой для эффективной аппаратурной реализации (возможно, аналоговым методом). Косинусное преобразование не производит точной диагонализации дискретного марковского процесса, однако экспериментальная проверка показала, что результат довольно близок к статистическому оптимуму Карунена — Лоэва [11].

Если марковская модель не дает удовлетворительного описания статистических характеристик процесса генерирования изображений, но предположение о стационарности остается в силе, то $[\Phi_x]$ и $[\Phi_y]$ сохраняют форму тёплицевых матриц. При больших матрицах отсчетов (больших N) и относительно быстром уменьшении корреляции между не слишком удаленными отсчетами изображения тёплицевы ковариационные матрицы могут быть хорошо аппроксимированы матрицами-циркулянтами, причем различия проявляются в конечных точках матриц:

$$[\Phi_x] \approx [C_x], \tag{2.33а}$$

$$[\Phi_y] \approx [C_y]. \tag{2.33б}$$

Такие «циркуляризованные» матрицы приводятся к диагональной форме с помощью дискретного преобразования Фурье, пример которого был представлен выше. Таким образом,

$$[C_x] = [\mathcal{F}][D_x][\mathcal{F}], \tag{2.34а}$$

$$[C_y] = [\mathcal{F}][D_y][\mathcal{F}], \tag{2.34б}$$

где

$$[\mathcal{F}] = \begin{matrix} x \downarrow \\ \end{matrix} \overset{u}{\longrightarrow} \left[\exp\left(\frac{2\pi i u x}{N}\right)\right], \tag{2.35}$$

а D_x и D_y — диагональные матрицы, ненулевые элементы которых (собственные значения) определяются разложением в ряд Фурье первой строки циркулянта [2.7]. Таким образом, область двумерного преобразования Фурье может представлять интерес для кодирования изображений, поскольку в этой области стационарные тёплицевы процессы оказываются почти некоррелированными. В какой степени преобразование Фурье пригодно для декорреляции данных, определяется тем, как велико число

N и насколько [Ф] приближается к диагональной матрице. Очевидное преимущество применения преобразования Фурье связано с наличием быстрого детерминированного алгоритма ($2N^2 \log_2 N$ операций вместо $2N^3$), хотя для его выполнения требуется сложная арифметика.

В строках 5—8 табл. 2.1 представлены наклонное преобразование, преобразование в дискретном линейном базисе, преобразования Уолша и Хаара. Каждое из этих преобразований имеет соответствующие статистические характеристики, описываемые разделимыми ковариационными матрицами вида

$$[\Phi_x^U] = [U][\Phi_x][U]^t, \tag{2.36а}$$

$$[\Phi_y^U] = [U][\Phi_y][U]^t, \tag{2.36б}$$

где $[U]$ означает соответственно [S], [DLB], [W] и [$\mathscr{H}$] [1]). Если принять сумму абсолютных величин недиагональных элементов в качестве меры неэффективности декорреляции процесса, выполняемой данным преобразованием, то наилучшим из четырех обсуждаемых преобразований окажется, по-видимому, наклонное преобразование, а наихудшим — преобразование Хаара. Если исходить из вычислительной точки зрения, то три преобразования — наклонное, в дискретном линейном базисе и Уолша — требуют выполнения порядка $2N^2 \log_2 N$ операций, причем под операциями понимают соответственно перемножения с плавающей запятой, перемножения целых чисел и сложения. Для преобразования Хаара требуется лишь $2(N-1)$ операций, однако оно не слишком хорошо декоррелирует данные. Преобразования Уолша и Хаара, примеры которых были приведены выше, требуют для их выполнения только операций сложения и вычитания и поэтому весьма привлекательны с точки зрения эффективной аппаратурной реализации. Правда, исходя из этой точки зрения, можно прийти к выводу, что преобразование тождественности еще более привлекательно, поскольку совсем не требует операций, но оно вообще не производит никакой декорреляции. Следовательно, компромисс между корреляцией и эффективностью реализации определяется на основе субъективной оценки [2]).

Следующая схема кодирования (строка 9 в табл. 2.1) часто называется гибридной, поскольку в ней используются методы

[1]) Матрицы обозначены по первым буквам названия преобразований: S (slant) — наклонное, W (Walsh) — Уолша, $\mathscr{H}$ (Haar) — Хаара, DLB (discrete linear basis) — в дискретном линейном базисе. — *Прим. перев.*

[2]) Пирл и др. [17] предложили для оценки преобразований метод на основе теории ε-энтропии, однако для аналитических решений считались практически приемлемыми только гауссовы источники. Это допущение вызывает сомнения, поскольку требует наличия бесконечных «хвостов» плотности распределения.

кодирования как в исходной, так и в преобразованной областях. В частности, изображение сначала подвергают одномерному унитарному преобразованию (косинусному, Фурье, Уолша и т. д.) в направлении строк, а затем полученные одномерные коэффициенты преобразования кодируют методом ДИКМ по преобразованным строкам. Обычно гибридные методы применяют для увеличения скорости работы, а также чтобы добиться уменьшения объема оборудования и требуемой емкости памяти. Благодаря применению таких методов становятся осуществимыми в реальном масштабе времени системы сжатия телевизионного спектра, причем требуется лишь nN ячеек памяти (здесь n — порядок предсказателя). Разумеется, требуемое число вычислительных операций сокращается до $N^2 \log_2 N + nN^2$, тогда как для быстрого двумерного преобразования оно составляет $2N^2 \log_2 N$.

В последней строке табл. 2.1 помещен чисто пространственный метод двумерной ДИКМ, рассмотренный в предыдущем разделе. Для системы двумерной ДИКМ требуется лишь n^2 ячеек памяти и $2nN^2$ вычислительных операций. Очевидно, что при $n = \log_2 N$ число требуемых операций для ДИКМ становится таким же, как для обычных методов быстрого преобразования. При $n = N$ получаем случай предсказания по полной матрице ($2N^3$), и число операций становится таким же, как при выполнении полномасштабного преобразования Карунена — Лоэва. Таким образом, преимущество метода ДИКМ состоит в наличии предсказателей низкого порядка (с малым n), которые производят эффективную декорреляцию данных и в то же время обеспечивают экономию в количестве вычислений и емкости памяти.

Выше были рассмотрены возможности использования различных преобразованных областей для получения менее коррелированных источников кодируемой информации, обеспечивающих сокращение полосы частот при передаче изображений. Практические алгоритмы кодирования не обсуждались, поскольку это выходит за рамки данной главы; заинтересованные читатели могут обратиться к работам [5, 8—17]. Пришлось также исключить различные дополнительные вопросы, связанные с кодированием изображения. Могло бы представить интерес наблюдение кодера при различных условиях работы канала, чтобы исследовать его характеристики при наличии шума. С другой стороны, на характеристики системы могут оказывать влияние дополнительные критерии, учитывающие аппаратурные ограничения и неидеальность синхронизации. Наконец, при проектировании системы кодирования изображения должен учитываться конечный получатель, т. е. наблюдатель с присущими ему параметрами.

2.4. Реставрация оригинала

Модель линейной системы отображения общего вида представлена во введении к этой главе (фиг. 2.1), где даны обозначения, используемые при последующем изложении. Обращаясь к дискретно-дискретному варианту модели и произведя «векторизацию» оригинала $\boldsymbol{f}$ и изображения $\boldsymbol{g}$ с помощью пакетного оператора [6, 18], мы получим (2.6)

$$\boldsymbol{g} = [H]\boldsymbol{f},$$

чему в непрерывно-непрерывной модели соответствует (2.1)

$$g(x, y) = \int_{-\infty}^{\infty}\int f(\xi, \eta)\, h(x, y, \xi, \eta)\, d\xi\, d\eta.$$

Смысл реставрации оригинала состоит в попытке произвести инверсию искажений, внесенных в оригинал $f(\xi, \eta)$ системой отображения. Решение системы, описываемой предыдущим уравнением, исследовалось для одного измерения Хансоном [19] и Вара [20], которые применили РСЗ в качестве средства получения псевдоинверсии уравнения. Исследования проводились в дискретном пространстве с использованием квадратурных формул интегрирования, чтобы получить эквивалентное представление в виде векторного пространства. Сондхи [21] предложил применить РСЗ как инструмент для обработки при реставрации изображений, подвергнутых пространственно-инвариантным искажениям. Трейтель и Шенкс [22] предложили плоскостные фильтры, которые проектируются с использованием метода РСЗ, подобного описанному в этой главе.

Введем четыре предположения, касающиеся импульсной характеристики $h(x, y, \xi, \eta)$, и затем оценим воздействие этих предположений на двумерные преобразования и инверсию связанных с ними искажающих явлений. Перечислим эти предположения в порядке возрастания сложности:

а) Разделимая пространственно-инвариантная функция рассеяния точки (РПИФРТ)

$$h(x, y, \xi, \eta) = a(x - \xi)\, b(y - \eta). \tag{2.37}$$

б) Разделимая пространственно-зависимая функция рассеяния точки (РПЗФРТ)

$$h(x, y, \xi, \eta) = a(x, \xi)\, b(y, \eta). \tag{2.38}$$

в) Неразделимая пространственно-инвариантная функция рассеяния точки (НПИФРТ)

$$h(x, y, \xi, \eta) = h(x - \xi, y - \eta). \tag{2.39}$$

г) Неразделимая пространственно-зависимая функция рассеяния точки (НПЗФРТ)

$$h(x, y, \xi, \eta) = h(x, y, \xi, \eta). \tag{2.40}$$

В качестве инструмента для выполнения процессов реставрации в следующих разделах используется алгебра матриц и внешних произведений, причем к матрице функции рассеяния точки $[H]$ в (2.6) не предъявляется каких-либо особых требований. Будем исходить из того, что известно аналитическое выражение $[H]$ и требуется восстановить f. При практическом применении модели удобно, чтобы система отображения сохраняла неизменной энергию изображения. Но поскольку скалярные элементы f и g сами являются некоторыми мерами энергии, имеем

$$\sum_{i=1}^{N^2} f_i = \sum_{i=1}^{N^2} g_i. \tag{2.41}$$

С другой стороны, импульс, или дельта-функция Дирака (точечный источник света), расположенный где-либо в f, вносит в g количество энергии, независимое от его положения в f. Следовательно, мы описываем систему с единичным усилением:

$$\sum_{i=1}^{N^2} h_{ij} = 1, \quad j = 1, \ldots, N^2. \tag{2.42}$$

Кроме того, природа чувствительных к энергии приборов такова, что они образуют только неотрицательные величины, и мы имеем

$$\begin{aligned} f_i &\geqslant 0, \\ g_i &\geqslant 0, \\ h_{ij} &\geqslant 0. \end{aligned} \tag{2.43}$$

Поскольку матрица $[H]$ неотрицательна, известно [23], что будет поэлементно неотрицательным следующее выражение:

$$[H_1] = \lambda_1^{1/2} \boldsymbol{u}_1 \boldsymbol{v}_1^t, \tag{2.44}$$

где λ_1, $\boldsymbol{u}_1$, $\boldsymbol{v}_1^t$ берутся из множества $\{\lambda_1, \boldsymbol{u}_i, \boldsymbol{v}_i\}$, определяемого с помощью РСЗ матрицы $[H]$.

2.4.1. Разделимая пространственно-инвариантная функция рассеяния точки

Этот случай отображения представляет собой, по-видимому, хорошее приближение первого порядка для правильно скорректированных линейных систем. Пространственно-инвариантная функция рассеяния точки подразумевает, что функциональная

форма импульсной характеристики не зависит от расположения точечного источника света (или импульсной характеристики) в плоскости исходного изображения. Предположение о разделимости подразумевает разделимость импульсной характеристики по прямоугольным координатам в виде (2.37)

$$h(x, y, \xi, \eta) = a(x - \xi)\, b(y - \eta).$$

В векторных обозначениях, соответствующих (2.6), получаем

$$[H] = [A] \otimes [B]^t, \tag{2.45}$$

где $[A]$ и $[B]$ — тёплицевы матрицы, а $\cdot\otimes$ — знак прямого, или кронекерова, произведения матриц. Благодаря прямоугольной разделимости импульсной характеристики мы можем использовать матричную форму записи без привлечения пакетного оператора:

$$[G] = [A][F][B], \tag{2.46}$$

где $[A]$ определяет смазывание столбцов, а $[B]$ — строк оригинала $[F]$. Поскольку матрица $[F]$ может быть неквадратной, а также поскольку матрицы смазывания могут оказаться сингулярными и вызывать невосполнимую потерю некоторых проявлений оригинала, инверсия искажений переходит в их псевдоинверсию. Однако если мы произведем «циркуляризацию» задачи, т. е. превратим матрицы смазывания в циркулянты (что эквивалентно периодическому повторению $[F]$ и $[G]$ в соответствующих плоскостях), то преобразование Фурье (2.35) обеспечит диагонализацию матриц смазывания. Это приводит к циклической свертке, и можно записать

$$[A] = [\mathcal{F}][\Lambda_a^{1/2}][\mathcal{F}], \tag{2.47}$$

$$[B] = [\mathcal{F}][\Lambda_b^{1/2}][\mathcal{F}]. \tag{2.48}$$

Подставляя эти выражения в (2.46), получаем

$$[G] = [\mathcal{F}][\Lambda_a^{1/2}][\mathcal{F}][F][\mathcal{F}][\Lambda_b^{1/2}][\mathcal{F}]. \tag{2.49}$$

Поскольку матрица $[\mathcal{F}]$ унитарна, а матрица $[\Lambda]$ диагональна, оригинал можно выразить в виде

$$[F] = [\mathcal{F}]^*[\Lambda_a^{-1/2}][\mathcal{F}]^*[G][\mathcal{F}]^*[\Lambda_b^{-1/2}][\mathcal{F}]^*, \tag{2.50а}$$

где звездочка * обозначает комплексное сопряжение. Однако внутренняя часть этого выражения представляет собой просто двумерное преобразование Фурье изображения $[G]$, взвешенное посредством обратных значений коэффициентов Фурье матрицы смазывания. Следовательно,

$$[F] = [\mathcal{F}]^*[\alpha][\mathcal{F}]^*, \tag{2.50б}$$

где

$$\alpha_{ij} = \lambda_{ai}^{-1/2}\lambda_{bi}^{-1/2}\boldsymbol{u}_i^t[G]\boldsymbol{u}_j, \tag{2.51}$$

а $\boldsymbol{u}_i$ представляет собой i-й столбец матрицы преобразования Фурье, благодаря которому $\boldsymbol{u}_i^t[G]\boldsymbol{u}_j$ делается равным ij-му коэффициенту Фурье изображения. При этом матрица $[\alpha]$ может быть интерпретирована как инверсно фильтрованное преобразование Фурье изображения. Тогда исходное изображение (оригинал) получается в виде двумерного преобразования матрицы $[\alpha]$, как следует из (2.50б). Из приведенного анализа вытекает, что устранение разделимого пространственно-инвариантного смазывания обеспечивается традиционной инверсной фурье-фильтрацией.

2.4.2. Разделимая пространственно-зависимая функция рассеяния точки

В этом случае функция рассеяния точки изменяет свою форму в процессе зондирования плоскости исходного изображения, оставаясь разделимой по прямоугольным координатам. Примерами подобных систем отображения могут служить радиолокаторы бокового обзора и рентгеновские аппараты с прямоугольным антикатодом.

ПЗФРТ может быть представлена в разделенной форме (2.38)

$$h(x, y, \xi, \eta) = a(x, \xi)\, b(\eta, y),$$

причем в дискретной матричной форме это выражение принимает прежний вид (2.45)

$$[H] = [A] \otimes [B]^t,$$

где матрица $[H]$ имеет размер $N^2 \times N^2$, а матрицы $[A]$ и $[B]$ имеют размер $N \times N$, но не обязательно являются тёплицевыми, как в случае ПИФРТ. Приятное следствие возможности представить ПЗФРТ в разделимой форме состоит в упрощении вычислений. Теперь в матричной записи не требуется использовать обозначений лексикографического или пакетного типа. Изображение $[G]$, как и прежде, выражается в виде (2.46)

$$[G] = [A][F][B],$$

где оригинал $[F]$ обрабатывается столбцовым оператором $[A]$ и строчным оператором $[B]$ по отдельности. Модель, соответствующая этому уравнению, исходит из условия отсутствия шума и предполагает, что ранг изображения $[G]$ определяется наименьшим из рангов $[A]$, $[F]$ и $[B]$. Если требуется инвертировать пространственно-зависимое смазывание, вносимое $[A]$ и $[B]$, то $[F]$ может быть получено в виде

$$[F] = [A]^{-1}[G][B]^{-1} \qquad (2.51а)$$

в предположении, что инверсные матрицы существуют. Выполнение этого условия практически весьма маловероятно, поскольку большинство систем отображения вызывает невосполнимую потерю некоторых частей оригинала, которые, естественно, уже не могут быть восстановлены. В этом случае может оказаться желательным найти минимальную по норме оценку оригинала $[\hat{F}]$, произведя псевдоинверсию матриц смазывания так, чтобы получаемые матрицы по рангам соответствовали матрицам смазывания. Псевдоинверсную оценку можно выразить в виде

$$[\hat{F}] = [A]^+ [G] [B]^+, \tag{2.51б}$$

где $[A]^+$ и $[B]^+$ означают псевдоинверсии $[A]$ и $[B]$ соответственно. Псевдоинверсию матриц, по-видимому, наиболее легко объяснить, используя ее разложение по сингулярным значениям (РСЗ) в форме

$$[A] = [U_a] [\Lambda_a]^{1/2} [V_a]^t \tag{2.51в}$$

и

$$[B] = [U_b] [\Lambda_b]^{1/2} [V_b]^t. \tag{2.51г}$$

Тогда имеем

$$[A]^+ = [V_a] [\Lambda_a]^{-1/2} [U_a]^t \tag{2.51д}$$

и

$$[B]^+ = [V_b] [\Lambda_b]^{-1/2} [U_b]^t. \tag{2.51е}$$

Здесь $[\Lambda_a]^{-1/2}$ и $[\Lambda_b]^{-1/2}$ — диагональные матрицы, ненулевые элементы которых равны обратным значениям ненулевых элементов $[\Lambda_a]^{1/2}$ и $[\Lambda_b]^{1/2}$. Переходя к векторным обозначениям, получаем

$$[A] = \sum_{i=1}^{R_a} \lambda_{ai}^{1/2} \boldsymbol{u}_{ai} \boldsymbol{v}_{ai}^t, \tag{2.51ж}$$

$$[B] = \sum_{i=1}^{R_b} \lambda_{bi}^{1/2} \boldsymbol{u}_{bi} \boldsymbol{v}_{bi}^t \tag{2.51з}$$

и

$$[A]^+ = \sum_{i=1}^{R_a} \lambda_{ai}^{-1/2} \boldsymbol{v}_{ai} \boldsymbol{u}_{ai}^t, \tag{2.52а}$$

$$[B]^+ = \sum_{i=1}^{R_b} \lambda_{bi}^{-1/2} \boldsymbol{v}_{bi} \boldsymbol{u}_{bi}^t, \tag{2.52б}$$

причем R_a и R_b представляют собой ранги матриц смазывания $[A]$ и $[B]$ соответственно. Используя представления с помощью РСЗ и описанные выше псевдоинверсии, можно теперь представить оценку оригинала в виде

$$[\hat{F}] = [V_a] [\Lambda_a]^{-1/2} [U_a]^t [G] [V_b] [\Lambda_b]^{-1/2} [U_b]^t, \tag{2.53а}$$

или

$$[\hat{F}] = [V_a][\alpha][U_b]^t, \tag{2.53б}$$

т. е.

$$[\hat{F}] = \sum_{i}^{R_a} \sum_{j}^{R_b} \alpha_{ij} \boldsymbol{v}_{ai} \boldsymbol{u}_{bj}^t, \tag{2.53в}$$

где матрица $[\alpha]$ определяется как

$$[\alpha] = [\Lambda_a]^{-1/2} [U_a]^t [G] [V_b] [\Lambda_b]^{-1/2}. \tag{2.54}$$

Полезно исследовать ij-й элемент матрицы $[\alpha]$, равный

$$\alpha_{ij} = \lambda_{ai}^{-1/2} \lambda_{bj}^{-1/2} \boldsymbol{u}_{ai}^t [G] \boldsymbol{v}_{bj}. \tag{2.55}$$

Здесь можно заметить сходство со случаем преобразования Фурье для ПИФРТ, рассмотренным выше. Однако векторы $\boldsymbol{u}_i$ и $\boldsymbol{v}_j$ теперь являются уже не столбцами матрицы преобразования Фурье, а векторами базисной системы, определяемой самими матрицами смазывания.

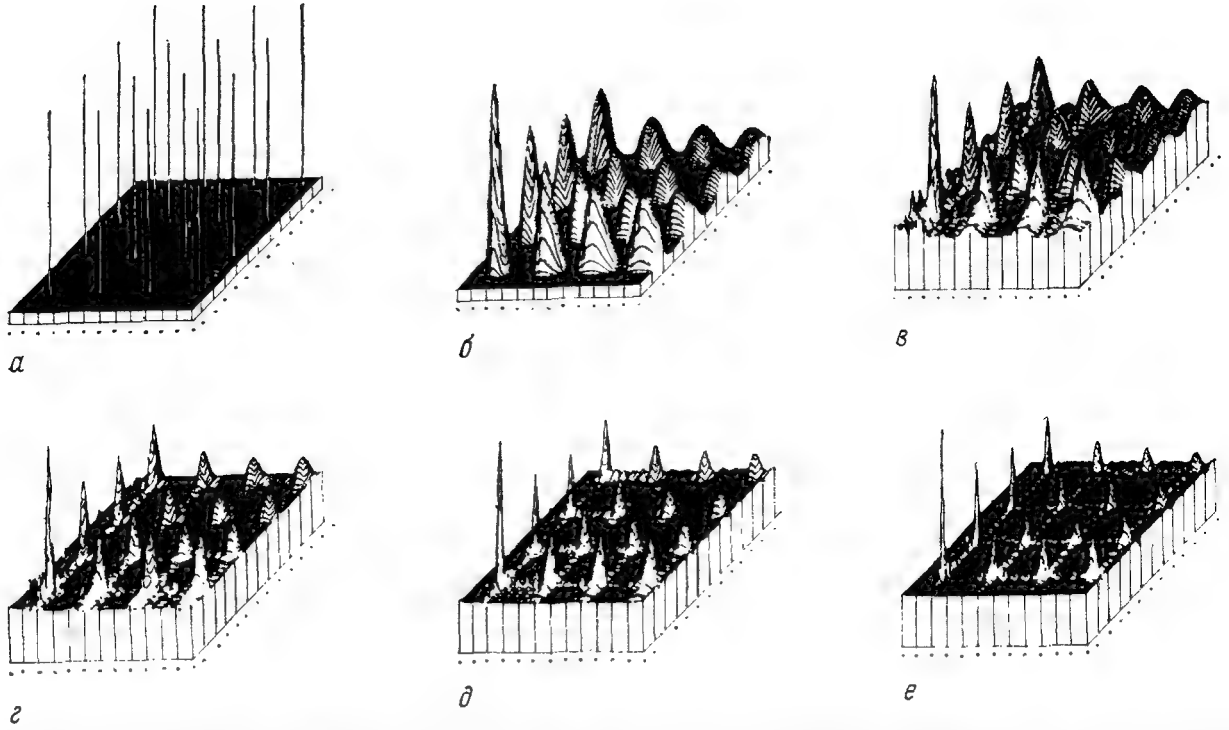
Вычисления для определения численных значений R_a и R_b связаны со значительными трудностями, обусловленными шумом машины и ошибками округления. Поэтому становится привлекательной усеченная оценка оригинала

$$[\hat{F}_K] = [V_a][\alpha_K][U_b]^t, \tag{2.56а}$$

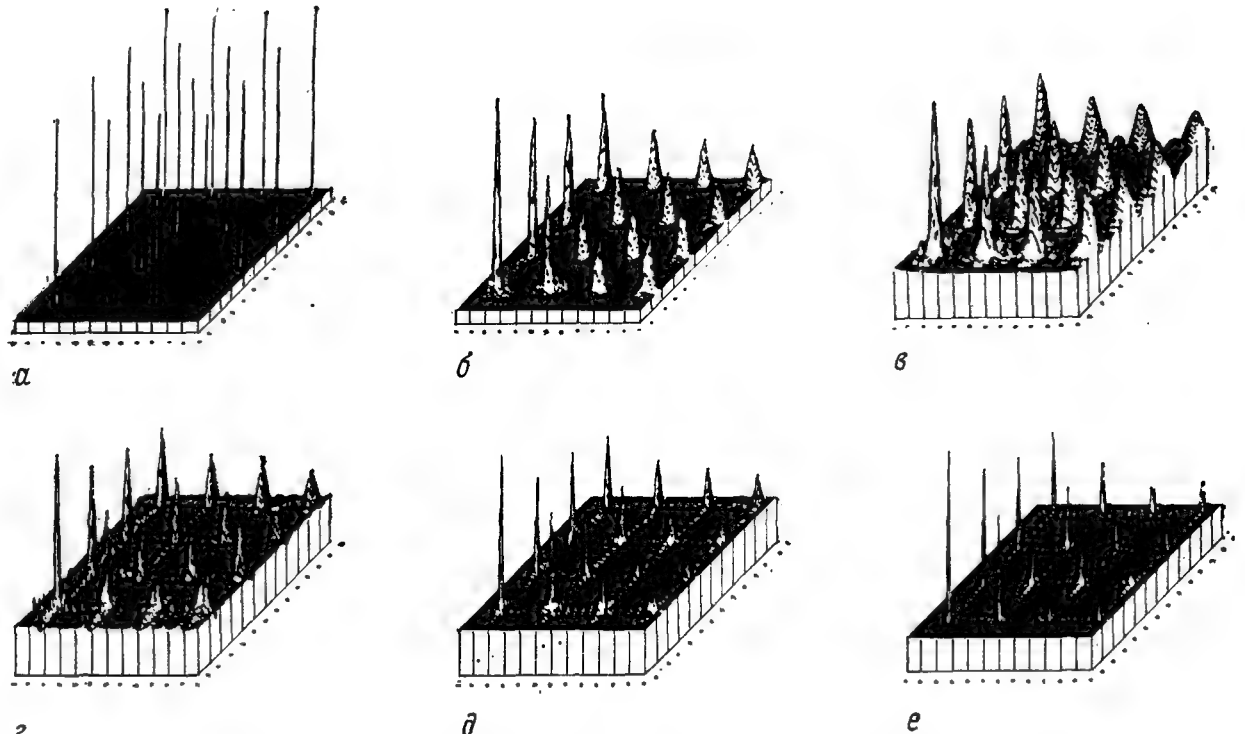
или

$$[\hat{F}_K] = \sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_{ij} \boldsymbol{v}_{ai} \boldsymbol{u}_{bj}^t. \tag{2.56б}$$

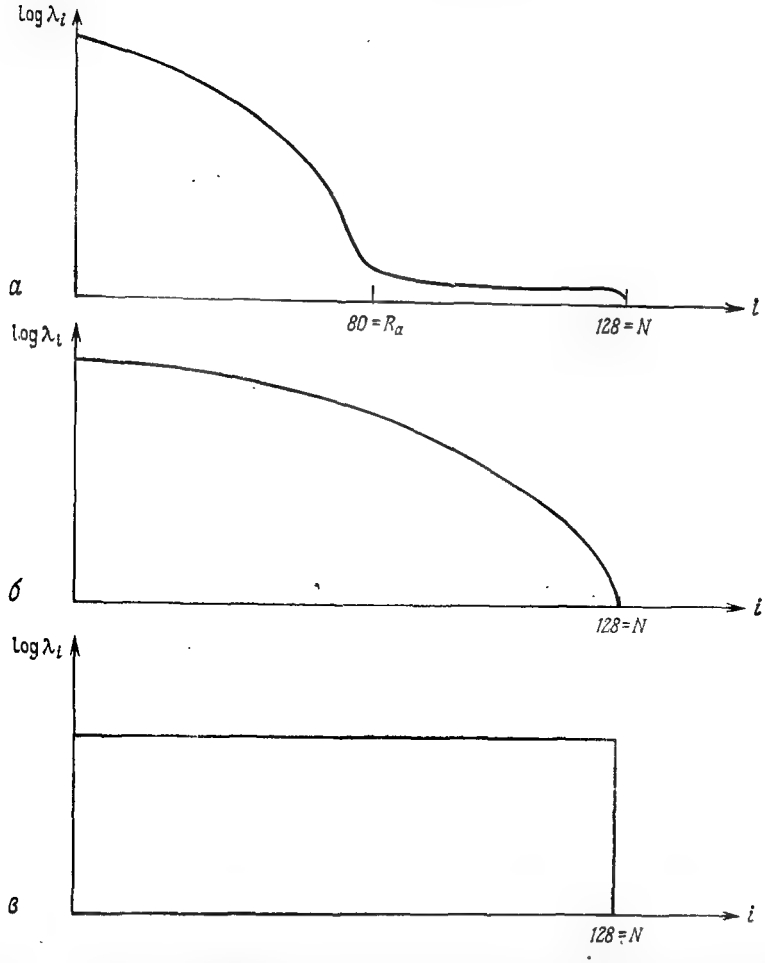
Как видно из этих выражений, усеченная оценка оригинала $[\hat{F}_K]$ представима в виде разложения в системе с двумерным базисом, определяемой столбцовыми и строчными ортогональными собственными векторами соответствующих матриц смазывания. Поскольку мы получаем псевдоинверсию, которая представляет собой оптимальную минимальную по норме среднеквадратичную оценку исходного изображения [2], разложение, определяемое пространствами $[V_a]$ и $[U_b]^t$, оказывается для такой инверсии наиболее эффективным. Чтобы несколько прояснить смысл этого утверждения, рассмотрим обсуждавшийся выше пример разделимой пространственно-инвариантной функции рассеяния точки (РПИФРТ), в котором матрицы $[A]$ и $[B]$ являются циркулянтами. В этом случае $[U_a] = [U_b] = [V_a] = [V_b] = [\mathcal{F}]$ (матрица дискретного преобразования Фурье) с элементами $\exp(2\pi i ux/N)$. Таким образом, $[\hat{F}_K]$ выражается посредством преобразования Фурье, при котором производится традиционная инверсная фильтрация.



Фиг. 2.15. Псевдоинверсия разделимой пространственно-зависимой функции рассеяния точки — случай сильных искажений: *а* — оригинал; *б* — искаженное изображение; *в* — псевдоинверсия, первые 21 собственное значение; *г* — псевдоинверсия, первые 31 собственное значение; *д* — псевдоинверсия, первые 41 собственное значение; *е* — псевдоинверсия, первые 51 собственное значение.

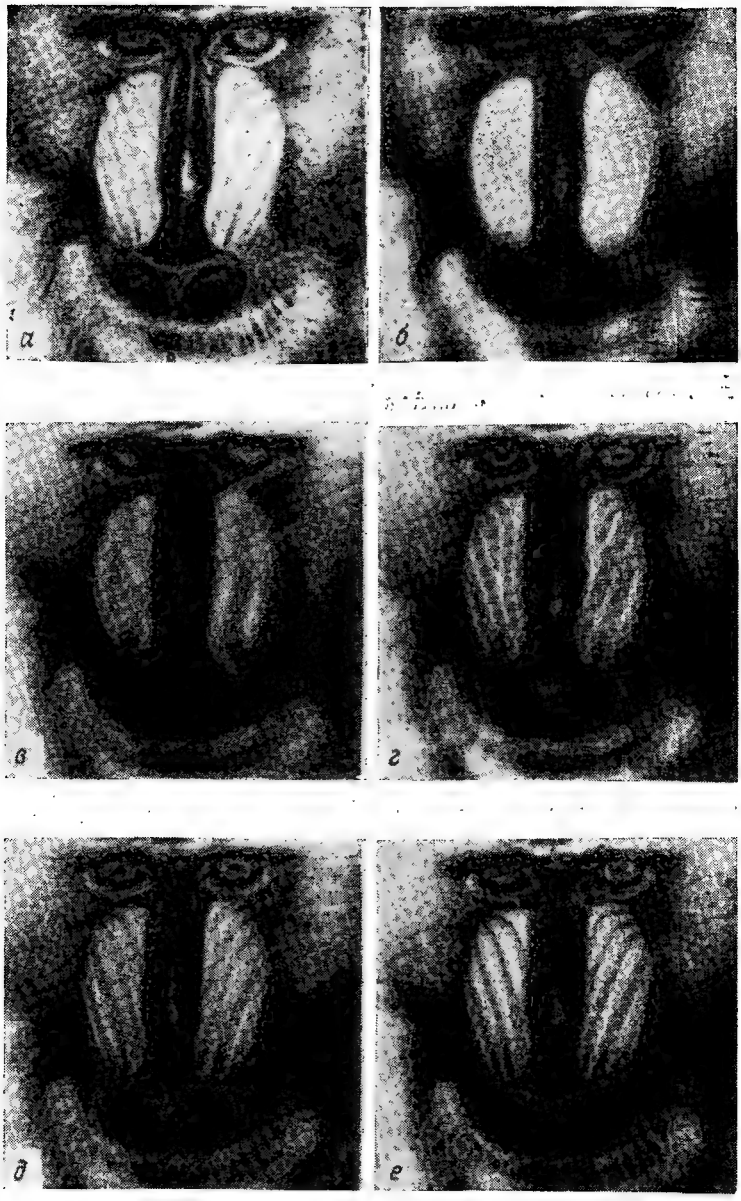


Фиг. 2.16. Псевдоинверсия разделимой пространственно-зависимой функции рассеяния точки — случай средних искажений: *а* — оригинал; *б* — искаженное изображение; *в* — псевдоинверсия, первые 21 собственное значение; *г* — псевдоинверсия, первые 41 собственное значение; *д* — псевдоинверсия, первые 61 собственное значение; *е* — псевдоинверсия, первые 81 собственное значение.



Фиг. 2.17. Графики изменения сингулярных значений при различных искажениях: *а* — сильные искажения; *б* — средние искажения; *в* — идеальное отображение (искажения отсутствуют).

Описанный выше анализ был произведен для случая отсутствия шума, при котором обобщенная инверсия (псевдоинверсия) дает в результате оптимальный фильтр для среднеквадратичной оценки. Обобщение на случай аддитивного шума при разделимой ПЗФРТ приводит к обобщенному псевдоинверсному винеровскому фильтру, причем энергетический спектр шума представляется в пространстве $[V_a]$, $[U_b]^t$, а не в двумерном фурье-пространстве.



Фиг. 2.18.Псевдоинверсия полутонового изображения, подвергнутого средним пространственно-зависимым искажениям: *а* — оригинал; *б* — искаженное изображение; *в* — псевдоинверсия, первые 21 собственное значение; *г* — псевдоинверсия, первые 41 собственное значение; *д* — псевдоинверсия, первые 61 собственное значение; *е* — псевдоинверсия, первые 81 собственное значение.

Изложенная теория была проверена путем моделирования с использованием матриц размера $N = 128$. Чтобы дать читателю лучшее представление о явлениях, происходящих в системе отображения с разделимой пространственно-зависимой функцией рассеяния точки, в качестве входного используется объект в виде 16 точечных источников света (фиг. 2.15 и 2.16). На приведенных иллюстрациях показан один из возможных вариантов разделимого пространственно-зависимого смазывания, при котором в левом нижнем углу изображения обеспечивается лучшая фокусировка, чем в правом верхнем. Методы Фурье в этом случае неприменимы вследствие очевидного пространственно-зависимого характера смазывания. При моделировании горизонтальное и вертикальное смазывания были сделаны одинаковыми. Смазанное изображение $[G]$ представлено на фиг. 2.15 и 2.16 с различной степенью искажений. Условный индекс (отношение наибольшего сингулярного значения к наименьшему), характеризующий искажения, составляет

$$C([A]) = 10^{16} \text{ для сильных искажений,}$$
$$C([A]) = 10^{12} \text{ для средних искажений,}$$
$$C([A]) = 1 \text{ для идеального случая,}$$

где под идеальным понимается случай отображения без искажений (т. е. $[A] = [B] = [I]$). На фиг. 2.17 представлены графики сингулярных значений для этих случаев. Возвращаясь к фиг. 2.15 и 2.16, мы видим, что псевдоинверсия смазанной решетки точечных источников обеспечивает реставрацию со все большей точностью по мере увеличения числа K. В случае сильных искажений псевдоинверсия явно вырождается при $K = 56$, при котором $C_{56}([A]) = 10^6$ и машинный шум становится преобладающим. В случае средних искажений $C_{56}([A]) = 20$, и инверсия остается стабильной даже после 81-й оценки, при которой $C_{81}([A]) = 10^3$.

Для проверки описанного метода реставрации при пространственно-зависимом смазывании и среднем уровне искажений была использована фотография павиана. Результаты представлены на фиг. 2.18, где видно, что средние искажения подавляют значительную часть деталей, имевшихся в исходном изображении, а псевдоинверсные оценки растущих порядков обеспечивают восстановление все большего количества деталей изображения.

2.4.3. Неразделимая пространственно-инвариантная функция рассеяния точки

Когда мы имеем неразделимую систему отображения (в прямоугольной или другой системе координат), объем вычислений существенно возрастает. В рассматриваемом случае

неразделимой ПИФРТ приходится использовать пакетный оператор. При этом матрица функции рассеяния точки $[H]$, соответствующая выражению (2.39)

$$h(x, y, \xi, \eta) = h(x - \xi, y - \eta),$$

принимает вид блочной тёплицевой матрицы размера $N^2 \times N^2$:

$$[H] = \begin{bmatrix} [A_1] & [A_2] \dots \\ [A_{-1}] & [A_1] & [A_2] \dots \\ [A_{-2}] & [A_{-1}] & [A_1] & [A_2] \dots \\ \vdots \\ \vdots \\ \cdot \end{bmatrix}, \tag{2.57}$$

где каждая отдельная $[A_i]$ также является тёплицевой матрицей. К сожалению, поскольку здесь отсутствует разделимость, мы не можем выразить $[H]$ через единственное кронекерово произведение. Однако если произвести «циркуляризацию» отдельных тёплицевых матриц, чтобы обеспечить выполнение циклической (а не полной) свертки, то $[H]$ преобразуется в блочный циркулянт, и это дает возможность произвести дополнительные упрощения [24, 25]. Коль скоро $[H]$ является блочной матрицей-циркулянтом, она диагонализируется матрицей $[\mathscr{F}_{N^2}]$, причем

$$[H] = [\mathscr{F}_{N^2}][D][\mathscr{F}_{N^2}]. \tag{2.58}$$

Здесь $[D]$ — диагональная, а $[\mathscr{F}_{N^2}]$ — блочная фурье-матрица (состоит из кронекерова произведения двух фурье-матриц размера $N \times N$). Таким образом,

$$[\mathscr{F}_{N^2}] = [\mathscr{F}] \otimes [\mathscr{F}]. \tag{2.59}$$

Между прочим, следует заметить, что столбцы или ряды матрицы $[\mathscr{F}_{N^2}]$ образуют обобщенные функции Уолша [26], и читателю следует учитывать, что $[\mathscr{F}_{N^2}]$ — это не фурье-матрица размера $N^2 \times N^2$, содержащая N^2 элементов в виде корней из единицы, а матрица $N^2 \times N^2$, содержащая N таких элементов. Таким образом, как следует из приведенного выше выражения, матрица $[\mathscr{F}_{N^2}]$ обеспечивает разделимость, поскольку это касается двумерного изображения. Иначе говоря, $[\mathscr{F}_{N^2}]$ представляет собой «пакетный» вариант двумерной матрицы преобразования Фурье. Однако она не обеспечивает разделимости функции рассеяния точки. Возвращаясь к проблеме реставрации, можем записать

$$\boldsymbol{g} = [H]\boldsymbol{f}, \tag{2.60а}$$

$$\boldsymbol{g} = [\mathscr{F}_{N^2}][D][\mathscr{F}_{N^2}]\boldsymbol{f}, \tag{2.60б}$$

но поскольку $[\mathcal{F}_{N^2}]$ — унитарная матрица (кронекерово произведение унитарных матриц унитарно), мы имеем

$$\hat{f} = [\mathcal{F}_{N^2}]^* [D]^+ [\mathcal{F}_{N^2}]^* \boldsymbol{g}, \tag{2.61}$$

где $[D]^+$ — диагональная матрица, элементы которой обратны элементам $[D]$. Однако мы не производим деления на нуль и поэтому элементы, соответствующие нулевым, заменяем нулями. Из этих рассуждений следует, что предыдущее уравнение может быть реализовано на ЭВМ, если произвести разделимое двумерное преобразование над $\boldsymbol{g}$ (т. е. выполнить $[\mathcal{F}_{N^2}]^* \boldsymbol{g}$, что соответствует пакетному преобразованию $[\mathcal{F}]^* [G] [\mathcal{F}]^*$), затем выполнить поточечное умножение на N^2 элементов $[D]^+$ и произвести разделимое двумерное преобразование над результирующим вектором $[D]^+ [\mathcal{F}_{N^2}] \boldsymbol{g}$. Описанная процедура, очевидно, есть не что иное, как типичная операция фильтрации, которая используется в огромном большинстве алгоритмов фильтрации, встречающихся при обработке изображений и выполняемых в области преобразования Фурье. Таким образом, двумерное преобразование Фурье, поточечное перемножение в преобразованной области и обратное преобразование Фурье векторно эквивалентны (2.61). Следовательно, при неразделимых ПИФРТ алгоритм реставрации может быть реализован как разделимый вплоть до операции фильтрации, выполняемой путем поточечного перемножения. Однако поскольку такое перемножение включает в себя только скалярные операции (а не векторные суммы), мы обходимся без перехода к пакетной системе обозначений в ЦВМ.

2.4.4. Неразделимая пространственно-зависимая функция рассеяния точки

Для систем отображения подобного типа о матрице функции рассеяния точки $h(x, y, \xi, \eta)$ можно сказать очень мало, если не имеется дополнительных аналитических сведений.

Интересные экстремальные условия представления $[H]$ с помощью РСЗ возникают в случае идеального отображения. При этом

$$[H] = [I],$$

и норма разности между $[H]$ и $[H_K]$ принимает значение

$$\| [H] - [H_K] \| = \sum_{i=K+1}^{N^2} \lambda_i.$$

Но поскольку все собственные значения $[H]$ равны единице, получаем

$$\| [H] - [H_K] \| = N^2 - K.$$

Заметим, что вследствие кратности собственных значений здесь не существует единственного разложения по внешним произведениям сингулярных векторов. К сожалению, получаемая аппроксимация оказывается весьма неудачной функцией K, и в предельном случае идеального отображения РСЗ фактически перестает быть привлекательным методом. В другом предельном случае, когда $[H]$ с точностью до масштабного коэффициента $(N^2)^{-1}$ определяется матрицей, целиком состоящей из единиц, мы имеем

$$h_{ij} = 1 \forall ij$$

и

$$\|[H] - [H_1]\| = 0,$$

поскольку

$$[H] = \mathbf{1}\mathbf{1}^t.$$

Другими словами, $[H]$ имеет ранг $R = 1$ и идеально представляется одним сингулярным значением и соответствующим ему внешним произведением. По-видимому, в промежутке между этими двумя предельными случаями существует некоторый континуум $[H]$.

При реставрации нам желательно инвертировать матрицу $[H]$ размера $N^2 \times N^2$ таким образом, чтобы получить наилучшую оценку оригинала $\hat{f}$. Следовательно,

$$\hat{\boldsymbol{f}} = [H]^+ \boldsymbol{g}, \tag{2.62}$$

и поскольку

$$[H]^+ = \sum_{i=1}^{R} \lambda_i^{-1/2} \boldsymbol{v}_i \boldsymbol{u}_i^t, \tag{2.63}$$

оригинал $\hat{f}$ выражается в виде

$$\hat{\boldsymbol{f}} = [H]^+ \boldsymbol{g} = \sum_{i=1}^{R} \lambda_i^{-1/2} (\boldsymbol{u}_i^t \boldsymbol{g}) \boldsymbol{v}_i. \tag{2.64}$$

Внутреннее произведение $(\boldsymbol{u}_i^t \boldsymbol{g})$ вместе с $\lambda_i^{-1/2}$ определяет скалярное взвешивание сингулярного вектора $\boldsymbol{v}_i$, и k-я оценка $\hat{\boldsymbol{f}}$ принимает вид

$$\hat{\boldsymbol{f}}_k = \hat{\boldsymbol{f}}_{k-1} + \lambda_k^{-1/2} (\boldsymbol{u}_k^t \boldsymbol{g}) \boldsymbol{v}_k, \tag{2.65}$$

где выражение для частичной суммы использовано, чтобы указать на сходимость к истинному оригиналу, если $[H]$ действительно несингулярна. Кроме того, вычисление частичных $\hat{\boldsymbol{f}}$ показывает, что в ЦВМ не обязательно одновременно хранить полный набор изображений и матриц функции рассеяния точки.

Без дополнительных сведений об $[H]$ инверсия этого наиболее общего случая систем отображения оказывается затруднительной. Однако, как было показано в предыдущих разделах, предположение о разделимости позволяет получить значительную экономию в вычислениях. Поэтому может оказаться полезной аппроксимация наиболее общей $[H]$ рядом разделимых функций. Таким образом,

$$h(x, y, \xi, \eta) \approx \sum_{i=1}^{I} a_i(x, \xi)\, b_i(y, \eta), \tag{2.66а}$$

или

$$[H] \approx \sum_{i=1}^{I} [A_i] \otimes [B_i]^t \tag{2.66б}$$

и, следовательно,

$$[G] \approx \sum_{i=1}^{I} [A_i][F][B_i]. \tag{2.67}$$

Однако инверсия $[H]$ есть инверсия суммы кронекеровых произведений, а не сумма их инверсий, и это порождает сомнения относительно полезности (2.66а). Возможна также полиномиальная аппроксимация степенным рядом в виде

$$h(x, y, \xi, \eta) \approx \sum_{i=1}^{I} c_i [A]^i \otimes [B]^{it}, \tag{2.68а}$$

или

$$[H] \approx \sum_{i=1}^{I} c_i [A]^i \otimes [B]^{it}, \tag{2.68б}$$

откуда следует

$$[G] \approx \sum_{i=1}^{I} c_i [A]^i [F] [B]^i. \tag{2.68в}$$

Однако и в этом случае требуется инвертировать сумму аппроксимирующего ряда. Таким образом, инверсия системы отображения наиболее общего вида связана с существенными вычислительными трудностями.

2.5. Заключение

В этой главе рассмотрены определенные аспекты обработки двумерных сигналов в приложении к обработке изображений на ЦВМ. Главной ее темой являются двумерные преобразования, их применения и реализации. Связующей нитью изложения служит матричная система обозначений на основе квадратичных форм, в соответствии с которой переход от изображения $[G]$

в двумерную преобразованную область осуществляется посредством матриц $[U]$ и $[V]$ в виде

$$[\alpha] = [U][G][V]^t.$$

Такие разделимые разложения изображений исследовались в связи с проблемой отображения как с математической точки зрения, так и с позиций кодирования изображений. Методы кодирования в пространственной и преобразованной областях проанализированы на основе символики двумерных преобразований, причем в первом случае использовались матрицы предсказания, а во втором — унитарные матрицы.

Затем методы двумерных преобразований рассматривались в приложении к решению задачи реставрации оригинала, для которой квадратичная форма имеет вид

$$[G] = [A][F][B],$$

где оригинал $[F]$ смазывается под действием $[A]$ и $[B]$, образуя изображение $[G]$. Инверсия явлений смазывания изображения рассмотрена для четырех общих случаев отображения, причем в тех случаях, когда предположение о разделимости не выполнялось, вводился оператор Кронекера, чтобы продемонстрировать читателям, каким образом предположение о разделимости можно применить к системам отображения наиболее общего вида.

В заключение автор выражает надежду, что читатели убедились в практической и методической ценности обобщенной матричной формулировки двумерных преобразований. Проще говоря, здесь показано, что методы матричной алгебры в соединении с линейной моделью систем отображения делают двумерные преобразования весьма мощным средством для выполнения многих видов обработки двумерных сигналов. Недостаток места не позволяет развить методы выполнения операций по выделению признаков, однако подход, аналогичный описанному, вполне пригоден для решения задач в этой области обработки изображений.

От автора

Данная глава не могла быть написана без неоценимой помощи ряда лиц и организаций. В целом автор хотел бы выразить благодарность за поддержку Институту обработки изображений (под руководством проф. У. Пратта) при Южно-Калифорнийском университете, Отделу оптических систем (под руководством А. Бордмана) фирмы «Эроспейс корпорейшн» и Группе обработки изображений (под руководством д-ров Д. Дженни и Б. Ханта) Лос-Аламосской научной лаборатории.

Настоящая работа финансировалась частично из внутренних фондов фирмы «Эроспейс корпорейшн» для субсидирования заказных исследований и разработок и частично Управлением перспективного планирования научно-исследовательских работ США по контракту № F 08606-72-C-0008 и контролировалась Восточным испытательным полигоном ВВС, база ВВС Патрик, штат Флорида.

Особую благодарность следует выразить проф. Хабиби из Южно-Калифорнийского университета, у которого заимствован подход к двумерным и гибридным системам ДИКМ как к способам перехода в преобразованную область. Программы разложения по сингулярным значениям, которые применялись для обработки изображений, были предоставлены К. Паттерсоном из фирмы «Эроспейс корпорейшн». Дополнительную информацию относительно РСЗ изображений заинтересованные читатели могут получить из статьи Паттерсона и Бюхлера [27]. Наконец, следует указать, что приведенные в этой главе фотографии предоставлены фирмой «Эроспейс корпорейшн».

Литература

1. *Golub G. H., Reinsch C.*, Numer. Math., **14**, 403 (1970).
2. *Albert A.*, Regression and the Moore-Penrose Pseudoinverse, Academic Press, New York, 1972.
3. *Andrews H. C.*, Computer Techniques in Image Processing, Academic Press, New York, 1970, Chap. 5. (Имеется перевод: *Эндрюс Г.* Применение вычислительных машин для обработки изображений. — М.: Энергия, 1977, гл. 5.)
4. *Harmuth H. F.*, Transmission of Information by Orthogonal Functions, 2nd ed., Springer, Berlin — Heidelberg — New York, 1972. (Имеется перевод 1-го изд.: *Хармут Х. Ф.* Передача информации ортогональными функциями. — М.: Связь, 1975.)
5. *Wintz P. A.*, Proc. IEEE, **60**, 809 (1972). [Имеется перевод: ТИИЭР, **60**, 69 (1972).]
6. *Ekstrom M. P.*, IEEE Trans. Computers, **C-22**, 322 (1973).
7. *Bellman R.*, Introduction to Matrix Analysis, McGraw-Hill Book Company, Inc., New York, 1960. (Имеется перевод: *Беллман Р.* Введение в теорию матриц. — М.: Наука, 1969.)
8. *Habibi A., Hershel R. S.*, IEEE Trans. Communication, **Com-22**, 692 (1974).
9. *Habibi A.*, Nat. Telecomm. Conf. Record, **1**, 12D-1 (November 1973).
10. *Huang T. S., Schrieber W. F., Tretiak O. J.*, Proc. IEEE, **59**, 1586 (1971). [Имеется перевод: ТИИЭР, **59**, 59 (1971).]
11. *Ahmed N., Natarayan T., Rao K. R.*, IEEE Trans. Computer, **C-23**, 90 (1974).
12. *Chen W.*, USCEE Rep. 441 (May 1973).
13. *Haralick R. M., Shanmugan K.*, IEEE Trans. Systems, Man and Cybernetics, **SMC-4**, 16 (1974).
14. *Pratt W. K., Kane J., Andrews H. C.*, Proc. IEEE, **57**, 58 (1969). [Имеется перевод: ТИИЭР, 57, 66 (1969).]
15. *Habibi A.*, IEEE Trans. Communication, **Com-22**, 614 (1974).
16. *Ray W. D., Driver E. M.*, IEEE Trans. Inform. Theory, **IT-16**, 663 (1970).
17. *Pearl J., Andrews H. C., Pratt W. K.*, IEEE Trans. Communication, **Com-20**, 411 (1972).

18. *Nissen D.*, Econometrics, **36**, 603 (1968).
19. *Hanson R. J.*, SIAM J. Numer. Anal., 8, 616 (1971).
20. *Varah J. M.*, SIAM J. Numer. Anal., **10**, 257 (1973).
21. *Sondhi M. M.*, Proc. IEEE, **60**, 842 (1972). [Имеется перевод: ТИИЭР, **60**, 108 (1972).]
22. *Treitel S., Shanks J. L.*, IEEE Trans. Geoscience Electronics, **GE-9**, 10 (1971).
23. *Todd J.*, Survey of Numerical Analysis, McGraw-Hill, New York, 1958, Chapt. 8.
24. *Ekstrom M. P.*, IEEE Trans. Computer, **C-23**, 320 (1974).
25. *Hunt B. R.*, IEEE Trans. Computer, **C-22**, 805 (1973).
26. *Chrestenson H. E.*, Pacific J. Mathematics, 5, 17 (1955).
27. *Patterson C. L., Buechler G. B.*, Computer, 7, 46 (1974).

3. Двумерные нерекурсивные фильтры

Дж. Фиасконаро

Данная глава посвящена главным образом рассмотрению четырех методов проектирования двумерных нерекурсивных цифровых фильтров. Это следующие методы: применение функций окна, частотная дискретизация, прямое применение линейного программирования и новый алгоритм, разработанный автором. Все теоретические сведения, необходимые для понимания четырех перечисленных алгоритмических методов, содержатся в разделе 3.1. В этом разделе рассматриваются двумерные дискретные системы общего вида, некоторые аспекты теории линейной аппроксимации и применение линейного программирования к проектированию фильтров. Раздел 3.2 содержит подробное описание всех четырех методов и примеры фильтров, спроектированных двумя рассмотренными методами. В разделе 3.3 дано краткое содержание главы и сделаны некоторые выводы.

Первые три метода (т. е. применение функций окна, частотная дискретизация и прямое применение линейного программирования) являются непосредственными обобщениями соответствующих методов проектирования одномерных нерекурсивных цифровых фильтров [1—5]. Только третий из этих методов пригоден для проектирования фильтров, оптимальных в том смысле, что максимальное абсолютное значение (чебышевская норма) ошибки принимает минимальное значение. В число других методов проектирования оптимальных нерекурсивных цифровых фильтров входит второй алгоритм Ремеза [6], а также методы, разработанные Херрманом [7], Хофстеттером и др. [8] и Парксом и Макклелланом [9]. К сожалению, ни один из этих методов не обобщается просто на двумерный случай.

Подход, основанный на прямом применении линейного программирования, в двумерном случае оказывается не таким эффективным, каким он является в одномерном случае. Это объясняется тем, что при заданном числе неизвестных коэффициентов импульсной характеристики в первом случае требуется иметь намного больше точек с ограничениями, чем во втором случае. Разработка нового алгоритма проектирования оптимальных двумерных фильтров преследовала цель решить эту проблему.

Новый алгоритм итеративен по своей сути, причем линейное программирование используется в нем как метод оптимизации; однако следует указать, что в каждой итерации достаточно иметь лишь небольшое число точек с ограничениями.

Новый алгоритм превосходит метод прямого применения линейного программирования, поскольку: а) «хорошие» аппроксимации обычно вычисляются быстро, причем эти аппроксимации всегда можно улучшить просто путем вычисления дополнительных итераций; б) наилучшие аппроксимации удается получить даже при чрезвычайно высокой плотности множества точек; в) обеспечивается возможность расчленить объемную задачу линейного программирования на ряд небольших задач. К сожалению, новый алгоритм оказался не таким эффективным с вычислительной точки зрения, как ожидалось. Имевшиеся вычислительные средства (ИБМ 360/67) ограничивали его применение проектированием фильтров, импульсная характеристика которых содержала не более 50 независимых коэффициентов.

3.1. Теория

3.1.1. Двумерные дискретные системы

Двумерная дискретная система есть оператор $\mathscr{L}$, который отображает входной массив $f(n, m)$ в выходной массив $g(n, m) = \mathscr{L}\{f(n, m)\}$, где n и m принимают положительные и отрицательные целочисленные значения. Пусть $g_1 = \mathscr{L}\{f_1\}$ и $g_2 = \mathscr{L}\{f_2\}$; система называется линейной, если

$$cg_1 + g_2 = \mathscr{L}\{cf_1 + f_2\}$$

для произвольной константы c. Более того, если $g = \mathscr{L}\{f\}$ и

$$g(n - n_0, m - m_0) = \mathscr{L}\{f(n - n_0, m - m_0)\}$$

для произвольных целых чисел n_0 и m_0, то система является инвариантной по отношению к сдвигу.

Единичный импульс $u_0(n, m)$ определяется выражением

$$u_0(n, m) = \begin{cases} 1 \text{ при } n = 0, \ m = 0, \\ 0 \text{ при других значениях.} \end{cases}$$

Очевидно, что любой массив $f(n, m)$ можно записать в виде суммы произведений констант на смещенный единичный импульс:

$$f(n, m) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f_{kl} u_0(n - k, m - l),$$

где константа $f_{kl} = f(k, l)$. Итак, если

$$g(n, m) = \mathscr{L}\{f(n, m)\}, \text{ то}$$

$$g(n, m) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f_{kl} \mathscr{L}\{u_0(n-k, m-l)\}.$$

Импульсная характеристика системы $h(n, m)$ определяется выражением

$$h(n, m) = \mathscr{L}\{u_0(n, m)\}.$$

Поэтому

$$g(n, m) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f(k, l)\, h(n-k, m-l). \tag{3.1}$$

Таким образом, выходной массив системы получается путем свертки ее входного массива с импульсной характеристикой этой системы. Можно показать, что

$$g(n, m) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f(n-k, m-l)\, h(k, l). \tag{3.2}$$

Если $h(n, m)$ тождественно равна нулю за пределами некоторой ограниченной области, то соответствующий фильтр называется нерекурсивным. В этом случае бесконечные пределы суммирования в (3.1) и (3.2) могут быть заменены конечными пределами. При этом каждый отсчет массива на выходе нерекурсивного фильтра представляет собой конечную взвешенную сумму отсчетов входного массива.

Цифровой фильтр называют устойчивым, если выходной массив остается ограниченным при любом ограниченном входном массиве. Цифровой фильтр устойчив тогда и только тогда, когда

$$\sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} |h(n, m)| < \infty.$$

На практике такую проверку бывает очень трудно выполнить, если фильтры не являются нерекурсивными [10, 11]. Однако для нерекурсивных фильтров этой трудности не существует. В этом случае сумма всегда должна быть ограниченной, поскольку $h(n, m)$ имеет нулевое значение за пределами конечной области.

Пусть входной массив имеет вид

$$f(n, m) = e^{j(\mu n + \nu m)},$$

где $j = \sqrt{-1}$. Легко показать, что при этом

$$g(n, m) = e^{j(\mu n + \nu m)} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} h(k, l)\, e^{-j(\mu k + \nu l)}. \tag{3.3}$$

Таким образом, комплексная экспонента является собственной функцией системы. Двойную сумму в (3.3) можно трактовать как передаточную функцию $H(\mu, \nu)$; эта функция непрерывна и периодична по обеим осям μ и ν, причем период составляет 2π. Точно так же двойную сумму в (3.3) можно трактовать как разложение $H(\mu, \nu)$ в двумерный ряд Фурье. Следовательно, $h(k, l)$ можно выразить через $H(\mu, \nu)$ в виде

$$h(k, l) = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} H(\mu, \nu)\, e^{j(\mu k + \nu l)}\, d\mu\, d\nu.$$

Двумерное z-преобразование есть обобщение передаточной функции; z-преобразование произвольного массива $f(n, m)$ выражается в виде

$$F(z_1, z_2) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} f(n, m)\, z_1^{-n} z_2^{-m},$$

где z_1 и z_2 — комплексные переменные. Передаточная функция представляет собой особый случай z-преобразования при $z_1 = \exp(j\mu)$ и $z_2 = \exp(j\nu)$. z-Преобразование выходного массива цифрового фильтра есть произведение z-преобразования входного массива на z-преобразование импульсной характеристики, т. е.

$$G(z_1, z_2) = F(z_1, z_2)\, H(z_1, z_2).$$

Обратное z-преобразование $g(n, m)$ получают, последовательно подвергая $G(z_1, z_2)$ (одномерному) обратному z_1-преобразованию и обратному z_2-преобразованию [12].

Все представленные до сих пор теоретические сведения о двумерных дискретных системах получены путем прямого обобщения теории одномерных дискретных систем [13]. Однако в некоторых случаях такое обобщение невозможно. Если в одномерной системе z-преобразование выражается в виде рациональной функции, то тогда числитель и знаменатель можно факторизовать и определить полюсы и нули данного z-преобразования. Когда положение полюсов становится известным, можно выполнить обратное z-преобразование для каждой возможной области сходимости z-преобразования.

В случае двумерных систем возникают две трудности. Во-первых, в общем случае полином невозможно факторизовать по двум (или большему числу) переменным; кроме того, такой полином, как правило, тождественно равен нулю в некотором непрерывном интервале значений двух переменных, а не в ограниченном числе дискретных точек. Во-вторых, четырехмерность $z_1 z_2$-пространства делает невозможным графическое представление областей сходимости; графическое представление подобных

областей в z-плоскости для функций одной переменной не представляет никаких затруднений. Указанные две трудности не позволили разработать теорию многомерных z-преобразований настолько же глубоко, насколько разработана теория одномерных z-преобразований.

При пользовании цифровой вычислительной машиной предпочтительно иметь дело с массивами, принимающими нулевое значение за пределами определенной ограниченной области. Кроме того, желательно, чтобы преобразование какого-то массива представляло собой другой массив, а не непрерывную функцию. Эта проблема решается с помощью дискретного преобразования Фурье. Если $f(n, m)$ тождественно равна нулю при $n < 0$ и $n \geqslant N$ и при $m < 0$ и $m \geqslant N$, то дискретное преобразование Фурье (ДПФ) записывается как

$$F(k, l) = \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} f(n, m) W^{-(nk+ml)},$$

где $W = \exp[j(2\pi/N)]$. Следует отметить, что такое преобразование дает только N^2 независимых значений $F(k, l)$ и эти значения являются отсчетами z-преобразования функции $f(n, m)$, определенными через равные интервалы при условии $|z_1| = |z_2| = 1$. Обратное ДПФ определяется как

$$f(n, m) = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} F(k, l) W^{(nk+ml)}.$$

Далее, если $F(k, l)$ и $H(k, l)$ являются ДПФ $f(n, m)$ и $h(n, m)$ соответственно, то обратное ДПФ произведения $H(k, l)F(k, l)$ есть циклическая свертка $f(n, m)$ и $h(n, m)$, что является аналогией одномерного случая. Эту операцию можно свести к линейной свертке, если соответствующим образом дополнить массивы $f(n, m)$ и $h(n, m)$ нулевыми значениями, прежде чем производить над ними ДПФ.

Обычно двумерное ДПФ вычисляется с использованием алгоритма одномерного быстрого преобразования Фурье (БПФ). Эта возможность обусловлена тем, что ДПФ можно записать в виде

$$F(k, l) = \sum_{m=0}^{N-1} \left(\sum_{n=0}^{N-1} f(n, m) W^{-nk} \right) W^{-ml}.$$

Сумма в скобках есть одномерное ДПФ строк массива $f(n, m)$, преобразующее его в новый массив $f'(k, m)$, столбцы которого затем подвергаются другому одномерному ДПФ. В результате получается $F(k, l)$. В общем случае требуется вычислить $2N$

одномерных ДПФ. Однако часто оказывается возможным существенно уменьшить указанное число преобразований, если воспользоваться определенными свойствами преобразуемых массивов.

3.1.2. Теория аппроксимации

Частотная характеристика цифрового фильтра определяется передаточной функцией, которая определена в предыдущем разделе как

$$H(\mu, \nu) = \sum_{n=-\infty}^{N} \sum_{m=-\infty}^{N} h(n, m)\, e^{-j(\mu n + \nu m)}$$

и которая является z-преобразованием $h(n, m)$, вычисленным при условии $z_1 = \exp(j\mu)$ и $z_2 = \exp(j\nu)$. В случае нерекурсивного фильтра частотная характеристика может быть записана в виде

$$H(\mu, \nu) = \sum_{n=-N}^{N} \sum_{m=-N}^{N} h(n, m)\, e^{-j(\mu n + \nu m)}, \qquad (3.4)$$

где принимается, что $h(n, m)$ тождественно равна нулю вне квадрата, каждая сторона которого содержит $2N + 1$ отсчет, а центр совпадает с началом координат. В случае нерекурсивных фильтров задача проектирования сводится к определению коэффициентов импульсной характеристики $h(n, m)$, обеспечивающих получение требуемой частотной характеристики $H(\mu, \nu)$.

Если коэффициенты импульсной характеристики являются чисто действительными, можно записать

$$H(\mu, \nu) = H^*(-\mu, -\nu),$$

где звездочкой обозначена комплексно-сопряженная величина. Если одновременно считать, что и частотная характеристика является чисто действительной (при этом фильтр вносит нулевой фазовый сдвиг), то

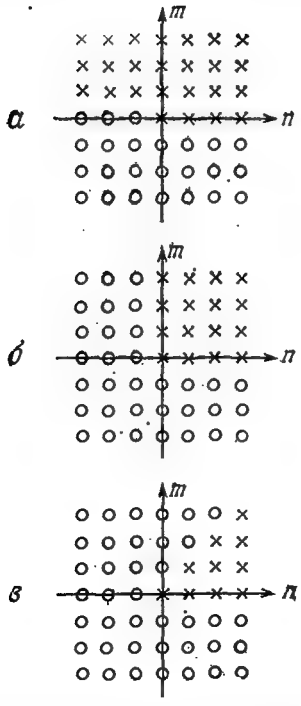
$$h(n, m) = h(-n, -m)$$

и вместо (3.4) можно записать

$$H(\mu, \nu) = \sum_{n=-N}^{\infty} \sum_{m=-N}^{\infty} h(n, m) \cos(n\mu + m\nu).$$

Поскольку $h(n, m) = h(-n, -m)$, это выражение можно записать в нескольких различных формах. Одна возможная форма записи имеет вид

$$H(\mu, \nu) = h(0, 0) + 2 \sum_{n=1}^{N} h(n, 0) \cos n\mu + \\ + 2 \sum_{m=1}^{N} \sum_{n=-N}^{N} h(n, m) \cos(n\mu + m\nu). \qquad (3.5)$$



Фиг. 3.1. Ограничения по симметрии; × — независимые коэффициенты импульсной характеристики; О — коэффициенты импульсной характеристики, обусловленные ограничениями по симметрии.

Указанную ситуацию иллюстрирует фиг. 3.1, *а* для случая импульсной характеристики, имеющей 7×7 отсчетов. Иногда желательно ввести дополнительные ограничения, касающиеся симметрии импульсной и частотной характеристик. Если принять, что

$$h(n, m) = h(|n|, |m|),$$

то

$$H(\mu, \nu) = H(|\mu|, |\nu|)$$

и частотную характеристику можно записать в виде (фиг. 3.1, *б*)

$$H(\mu, \nu) = h(0, 0) + 2\sum_{n=1}^{N} h(n, 0)\cos n\mu + 2\sum_{m=1}^{N} h(0, m)\cos m\nu + \\ + 2\sum_{m=1}^{N}\sum_{n=1}^{N} h(n, m)[\cos(n\mu + m\nu) + \cos(n\mu - m\nu)]. \quad (3.6)$$

Наконец, если принять, что

$$h(n, m) = h(m, n) \quad \text{и} \quad h(n, m) = h(|n|, |m|),$$

то

$$H(\mu, \nu) = H(\nu, \mu) \quad \text{и} \quad H(\mu, \nu) = H(|\mu|, |\nu|)$$

и частотную характеристику можно записать в виде (фиг. 3.1, *в*)

$$H(\mu, \nu) = h(0, 0) + 2\sum_{n=1}^{N} h(n, 0)[\cos n\mu + \cos n\nu] +$$
$$+ 2\sum_{n=1}^{N} h(n, n)[\cos(n\mu + n\nu) + \cos(n\mu - n\nu)] +$$
$$+ 2\sum_{m=1}^{N}\sum_{n=m+1}^{N} h(n, m)[\cos(n\mu + m\nu) +$$
$$+ \cos(n\mu - m\nu) + \cos(m\mu + n\nu) + \cos(m\mu - n\nu)]. \qquad (3.7)$$

Можно ли с помощью выражений (3.4)—(3.7) реализовать абсолютно точно «идеальные» фильтры? Если речь идет об идеальных фильтрах общего вида, то ответ на этот вопрос будет отрицательным. Например, для идеального «кругового» фильтра нижних частот частотная характеристика $H(\mu, \nu)$ имеет вид

$$H(\mu, \nu) = \begin{cases} 1 \text{ при } \sqrt{\mu^2 + \nu^2} \leqslant R, \\ 0 \text{ при других значениях} \end{cases}$$

в пределах одного периода, например при $-\pi \leqslant \mu \leqslant \pi$ и $-\pi \leqslant \nu \leqslant \pi$. Соответствующая импульсная характеристика имеет вид

$$h(n, m) = \frac{RJ_1(R\sqrt{n^2 + m^2})}{2\pi\sqrt{n^2 + m^2}},$$

где J_1 — функция Бесселя первого рода первого порядка. Аналогично частотная характеристика идеального кругового полосового фильтра имеет вид

$$H(\mu, \nu) = \begin{cases} 1 \text{ при } R_1 \leqslant \sqrt{\mu^2 + \nu^2} \leqslant R_2, \\ 0 \text{ при других значениях,} \end{cases}$$

а его импульсная характеристика имеет вид

$$h(n, m) = \frac{R_2 J_1(R_2\sqrt{n^2 + m^2})}{2\pi\sqrt{n^2 + m^2}} - \frac{R_1 J_1(R_1\sqrt{n^2 + m^2})}{2\pi\sqrt{n^2 + m^2}}.$$

Идеальный круговой фильтр верхних частот имеет частотную характеристику

$$H(\mu, \nu) = \begin{cases} 0 \text{ при } \mu^2 + \nu^2 \leqslant R^2, \\ 1 \text{ при других значениях} \end{cases}$$

и импульсную характеристику

$$h(n, m) = \begin{cases} 1 \text{ при } n = 0 \text{ и } m = 0, \\ -\dfrac{RJ_1(R\sqrt{n^2+m^2})}{2\pi\sqrt{n^2+m^2}} \text{ при других значениях.} \end{cases}$$

Идеальный двумерный дифференциатор одного из возможных типов обладает частотной характеристикой вида

$$H(\mu, \nu) = \mu^2 + \nu^2.$$

Соответствующая импульсная характеристика имеет вид

$$h(n, m) = \begin{cases} \dfrac{2\pi^2}{3} & \text{при } n = 0 \text{ и } m = 0, \\ \dfrac{2}{n^2}\cos n\pi & \text{при } n \neq 0 \text{ и } m = 0, \\ \dfrac{2}{m^2}\cos m\pi & \text{при } n = 0 \text{ и } m \neq 0, \\ 0 & \text{при других значениях.} \end{cases}$$

Во всех рассмотренных случаях импульсная характеристика имеет бесконечную протяженность, т. е. не существует конечной области, за пределами которой импульсная характеристика была бы тождественно равна нулю. Следовательно, рассмотренные фильтры не являются нерекурсивными и не могут быть реализованы с помощью выражений (3.4)—(3.7).

Поскольку частотные характеристики таких фильтров нельзя реализовать абсолютно точно, их следует аппроксимировать. При этом необходимо пользоваться некоторой мерой, характеризующей степень приближения аппроксимированной частотной характеристики к требуемой, некоторым критерием, позволяющим выбрать наиболее точную или наилучшую аппроксимацию. Единственный выбор в обоих случаях падает на чебышевскую норму [6]. Для некоторой функции F эта норма определяется как максимальное абсолютное значение F и записывается как $\|F\|$. Пусть $H(\mu, \nu)$ — аппроксимация идеальной частотной характеристики $F(\mu, \nu)$; тогда степень приближения $H(\mu, \nu)$ к $F(\mu, \nu)$ есть чебышевская норма функции ошибки, т. е.

$$\|\text{ошибка}\| = \max|F(\mu, \nu) - H(\mu, \nu)|.$$

Наилучшей аппроксимацией $F(\mu, \nu)$ называют такую $H(\mu, \nu)$, которая минимизирует чебышевскую норму функции ошибки.

Некоторые фундаментальные выводы теории линейной чебышевской аппроксимации применимы к задаче проектирования

фильтров в том виде, в каком она формулировалась выше. В частности, важное значение имеют три теоремы: теорема существования, теорема идентификации и теорема единственности. Эти теоремы приводятся здесь без доказательств, поскольку последние содержатся в литературе [6]. Все три теоремы сформулированы применительно к задаче проектирования двумерных фильтров. Сделано это исключительно ради удобства изложения. Все результаты, применимые к двумерному случаю, в равной степени применимы к трех-, четырех- или n-мерному случаю.

Теорема существования. Конечномерное линейное подпространство нормированного линейного пространства содержит по меньшей мере одну точку, расположенную на минимальном расстоянии от любой фиксированной точки.

Чтобы показать, каким образом эта теорема используется при решении задачи проектирования фильтров, необходимо отметить следующее: множество c всех функций, являющихся непрерывными на замкнутом подмножестве области $-\pi \leqslant \mu \leqslant \pi$, $-\pi \leqslant \nu \leqslant \pi$, представляет собой нормированное линейное пространство, если используется чебышевская норма и если сложение и скалярное умножение определяются соотношениями

$$(f+g)(\mu, \nu) = f(\mu, \nu) + g(\mu, \nu),$$
$$(\lambda f)(\mu, \nu) = \lambda f(\mu, \nu)$$

для любой действительной константы λ и для f и g как элементов c. Более того, выражения (3.5)—(3.7) определяют конечномерное линейное подмножество этого нормированного линейного пространства. Таким образом, если «идеальный» фильтр принадлежит нормированному линейному пространству, существует по меньшей мере одна аппроксимация, которая является наиболее близкой (в смысле чебышевской нормы) к этому «идеальному» фильтру.

Выражения (3.5)—(3.7) можно записать в форме обобщенных полиномов вида

$$H(\mu, \nu) = \sum_{i=1}^{n} c_i g_i(\mu, \nu),$$

где c_i — коэффициенты импульсной характеристики, $g_i(\mu, \nu)$ — различные косинусные функции. Используя такое представление и принимая, что $f(\mu, \nu)$ есть идеальная частотная характеристика, можно сформулировать теорему идентификации.

Теорема идентификации. Для того чтобы при выборе коэффициентов $c_1, \ldots, c_n$ чебышевская норма $r(\mu, \nu) = \sum c_i g_i(\mu, \nu) - f(\mu, \nu)$ [1]) имела минимальное значение, необходимо и доста-

[1]) Если пределы суммирования не указаны, следует подразумевать пределы 1 и n.

точно, чтобы начало координат n-мерного пространства было расположено на выпуклой оболочке множества точек

$$\{r(\mu, \nu)[g_1(\mu, \nu), \ldots, g_n(\mu, \nu)] : |r(\mu, \nu)| = \|r(\mu, \nu)\|\}.$$

Выпуклая оболочка множества A определяется как множество точек p, которые могут быть выражены конечными суммами вида $p = \sum \theta_i f_i$ при $f_i \in A$, $\theta_i \geqslant 0$ и $\sum \theta_i = 1$. Сформулированная иначе, теорема утверждает, что $\sum c_i g_i$ является наилучшей аппроксимацией, если и только если существуют, например, m точек, для которых $|r(\mu_i, \nu_i)| = \|r(\mu, \nu)\|$ и

$$\begin{bmatrix} g_1(\mu_1, \nu_1) & g_1(\mu_2, \nu_2) & \ldots & g_1(\mu_m, \nu_m) \\ g_2(\mu_1, \nu_1) & g_2(\mu_2, \nu_2) & \ldots & g_2(\mu_m, \nu_m) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ g_n(\mu_1, \nu_1) & g_n(\mu_2, \nu_2) & \ldots & g_n(\mu_m, \nu_m) \end{bmatrix} \begin{bmatrix} \theta_1 r(\mu_1, \nu_1) \\ \theta_2 r(\mu_2, \nu_2) \\ \cdot \\ \cdot \\ \cdot \\ \theta_m r(\mu_m, \nu_m) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix},$$

откуда вытекает, что ограничение $\sum \theta_i = 1$ можно заменить на $\sum \theta_i > 0$. Рассматриваемая теорема позволяет распознать, является или нет данная аппроксимация наилучшей. Однако она не дает данных для практического вычисления наилучшей аппроксимации. Эта проблема обсуждается дальше в этом разделе, а также в следующем разделе.

Теорема единственности (теорема Хаара). Наилучшая аппроксимация непрерывной функции f обобщенным полиномом $\sum c_i g_i$ является единственной при любом выборе f, если и только если $\{g_1, \ldots, g_n\}$ удовлетворяет условию Хаара.

Множество непрерывных функций $\{g_1, \ldots, g_n\}$ удовлетворяет условию Хаара, если каждый детерминант

$$D = \begin{vmatrix} g_1(\mu_1, \nu_1) & \ldots & g_n(\mu_1, \nu_1) \\ g_1(\mu_n, \nu_n) & \ldots & g_n(\mu_n, \nu_n) \end{vmatrix},$$

составленный из n отдельных точек, не равен нулю. Другая формулировка: система функций $\{g_1, \ldots, g_n\}$ удовлетворяет условию Хаара, если и только если не существует обобщенного полинома $\sum c_i g_i$, имеющего более чем $n - 1$ корней. Систему функций, удовлетворяющую условию Хаара, иногда называют чебышевской системой.

Сразу же возникает следующий вопрос. Удовлетворяют ли косинусные функции (3.5)—(3.7) условию Хаара в пределах области $-\pi \leqslant \mu \leqslant \pi$, $-\pi \leqslant \nu \leqslant \pi$? Нет, не удовлетворяют по следующим причинам [14]. Предположим, что были выбраны такие n точек, что детерминант D не равен нулю. Будем перемещать

две точки, например точки (μ_1, ν_1) и (μ_2, ν_2), таким образом, чтобы они всегда оставались в пределах области $-\pi \leqslant \leqslant \mu \leqslant \pi$, $-\pi \leqslant \nu \leqslant \pi$, чтобы всегда существовало n отдельных точек и чтобы точка (μ_1, ν_1) остановилась там, где начала свое перемещение точка (μ_2, ν_2), а точка (μ_2, ν_2) остановилась там, где начала свое перемещение точка (μ_1, ν_1). Совершенно очевидно, что при этом произойдет перестановка двух строк детерминанта D, т. е. детерминант изменит свой знак. Поскольку первоначально детерминант не был равен нулю, перемена знака при непрерывном изменении детерминанта означает обязательное его прохождение через нулевое значение. Следовательно, рассматриваемые функции не удовлетворяют условию Хаара. В литературе [15] имеется формальное доказательство этой теоремы. Показано, что в общем случае для функций двух и более переменных не существует нетривиальных чебышевских систем.

Когда множество функций $\{g_1, \ldots, g_n\}$ удовлетворяет условию Хаара, это не только означает, что наилучшая аппроксимация является единственной: появляется возможность доказать одну исключительно мощную теорему идентификации, известную под названием теоремы чередования [6]. Эта теорема применима только к задачам одномерной аппроксимации.

Теорема чередования. Пусть $\{g_1, \ldots, g_n\}$ — система функций, которые являются непрерывными на $[a, b]$ и удовлетворяют условию Хаара, и пусть X — любое замкнутое подмножество интервала $[a, b]$. Для того чтобы некоторый обобщенный полином $P = \sum c_i g_i$ являлся наилучшей аппроксимацией на X заданной функции f, являющейся непрерывной на X, необходимо и достаточно, чтобы функция ошибки $r = f - P$ претерпевала на X по меньшей мере $n + 1$ «уклонение» следующего вида: $r(x_i) = -r(x_{i-1}) = \pm \|r\|$, причем $x_0 < \ldots < x_n$ и $x_i \in X$.

Для задачи проектирования одномерного фильтра фундаментальное значение имеют два множества функций: $\{1, x, x^2, \ldots, x^n\}$ и $\{1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \cos nx, \sin nx\}$. Первое удовлетворяет условию Хаара на любом интервале, второе — на интервале $[0, 2\pi]$ (см. [16]). Благодаря этому проектировать одномерные фильтры значительно проще, чем двумерные.

Задачу проектирования двумерного фильтра можно сформулировать несколько иначе [17], если начать с системы (возможно, бесконечной степени) несовместных уравнений вида

$$F(\mu, \nu) = \sum_{i=1}^{n} c_i g_i(\mu, \nu),$$

где $F(\mu, \nu)$ — идеальная (нереализуемая) частотная характеристика. Задача состоит в том, чтобы найти множество $\{c_1, \ldots, c_n\}$,

которое минимизирует чебышевскую норму функции Δ, определяемой как

$$\Delta = F(\mu, \nu) - \sum_{i=1}^{n} c_i g_i(\mu, \nu).$$

Конечномерная подсистема

$$\Delta_m = F(\mu_m, \nu_m) - \sum_{i=1}^{n} c_i g_i(\mu_m, \nu_m), \quad m = 1, \ldots, M,$$

исходной системы уравнений называется ограничивающей, если все Δ_m не равны нулю, а абсолютные значения всех Δ_m не могут быть одновременно уменьшены при любом выборе множества $\{c_1, \ldots, c_n\}$. Чтобы исследовать изменения абсолютных значений Δ_m в зависимости от изменения множества $\{c_1, \ldots, c_n\}$, полезно записать подсистему уравнений в виде

$$\operatorname{sign}\Delta_m \cdot \Delta_m = \operatorname{sign}\Delta_m \cdot F(\mu_m, \nu_m) - \operatorname{sign}\Delta_m \sum_{i=1}^{n} (c_i + d_i) g_i(\mu_m, \nu_m),$$

где d_i выражает изменение c_i. Теперь левая часть каждого уравнения равна абсолютному значению Δ_m. Если система неравенств

$$\operatorname{sign}\Delta_m \cdot \sum_{i=1}^{n} d_i g_i(\mu_m, \nu_m) > 0, \quad m = 1, \ldots, M,$$

является несовместной, т. е. если невозможно удовлетворить всем этим неравенствам при произвольном выборе множества $\{d_1, \ldots, d_n\}$, то ни при каких условиях невозможно одновременно уменьшить абсолютные значения всех Δ_m. Можно показать (с помощью теоремы о линейных неравенствах [6]), что эта система неравенств является несовместной, если и только если начало координат n-мерного пространства расположено на выпуклой оболочке множества $\{\operatorname{sign}\Delta_m [g_1(\mu_m, \nu_m), \ldots, g_n(\mu_m, \nu_m)]\}$. Эти предварительные замечания позволяют доказать следующую теорему, являющуюся вариантом обобщенной теоремы Валле — Пуссена.

Теорема. Если P есть обобщенный полином, т. е. $P(\mu, \nu) = \sum c_i g_i(\mu, \nu)$, а X есть конечномерное множество точек, для которого система уравнений

$$\Delta_m = F(\mu_m, \nu_m) - \sum_{i=1}^{n} c_i g_i(\mu_m, \nu_m), \quad m = 1, \ldots, M,$$

является ограничивающей, то

$$\min_X |P - F| \leqslant \max_X |P^* - F| \leqslant \max_X |P - F|,$$

где P^* — наилучшая чебышевская аппроксимация F на множестве X.

Доказательство. Правое неравенство является тривиальным, поскольку P есть возможное наилучшее приближение к F на X. Чтобы доказать левое неравенство, допустим, что

$$\max_X |P^* - F| < \min_X |P - F|.$$

Это означает, что все значения $|P - F|$ были одновременно уменьшены в процессе перехода от P к P^*. Однако это противоречит тому условию, что система

$$\Delta_m = F(\mu_m, \nu_m) - \sum_{i=1}^{n} c_i g_i(\mu_m, \nu_m), \quad m = 1, \ldots, M,$$

является ограничивающей. Следовательно, левое неравенство доказано.

Обобщенная теорема Валле — Пуссена лежит в основе алгоритма, с помощью которого можно вычислить наилучшую чебышевскую аппроксимацию $P^*(\mu, \nu)$ идеальной частотной характеристики $F(\mu, \nu)$ на множестве (возможно, бесконечномерном) точек X. Этот алгоритм можно описать следующим образом. Пусть P_k — наилучшая аппроксимация F на $X_k \subseteqq X$, причем

$$P_k(\mu, \nu) = \sum_{i=1}^{n} c_i^k g_i(\mu, \nu),$$

$$\|E_k\| = \max_X |P_k - F|,$$

$$\delta_k = \max_{X_k} |P_k - F|.$$

Далее определяется новое множество точек $X_{k+1} \subseteqq X$, такое, что

$$\max_{X_{k+1}} |P_k - F| = \|E_k\|,$$

$$\min_{X_{k+1}} |P_k - F| \geqslant \delta_k,$$

и такое, что система уравнений

$$\Delta = F(\mu, \nu) - \sum_{i=1}^{n} c_i g_i(\mu, \nu) \text{ для всех } (\mu, \nu) \in X_{k+1}$$

является ограничивающей. Согласно теореме, имеем

$$\min_{X_{k+1}} |P_k - F| \leqslant \max_{X_{k+1}} |P_{k+1} - F| \leqslant \max_{X_{k+1}} |P_k - F|,$$

или

$$\delta_k \leqslant \delta_{k+1} \leqslant \|E_k\|, \tag{3.8}$$

где P_{k+1} — наилучшая аппроксимация F на X_{k+1} и

$$\delta_{k+1} = \max_{X_{k+1}} |P_{k+1} - F|.$$

Это означает, что δ должна возрастать (или оставаться неизменной — условие, никогда не наблюдаемое практически) при переходе от итерации k к итерации $k+1$. Алгоритм должен сходиться, поскольку наилучшая аппроксимация P^* существует и

$$\max_{X} |P^* - F| = \delta^* < \infty.$$

Точный метод определения множества X_{k+1} по P_k и X_k станет более понятным после рассмотрения линейного программирования.

Алгоритмы, подобные рассмотренному, можно подразделить на три широкие категории [18]: минимизирующие методы, максимизирующие методы и минимаксные методы. Любой минимизирующий метод гарантирует понижение верхней границы ошибки от итерации к итерации. Соответственно максимизирующий метод гарантирует повышение нижней границы ошибки. Наконец, минимаксный метод обеспечивает одновременно повышение нижней и понижение верхней границ ошибки. Неравенство (3.8) ясно показывает, что рассмотренный выше алгоритм является максимизирующим методом, поскольку δ_k есть нижняя граница ошибки δ^*, причем гарантируется повышение δ_k при переходе от итерации k к итерации $k+1$; однако не гарантируется, что верхняя граница ошибки $\|E_k\|$ понижается.

3.1.3. Линейное программирование

Для решения задачи проектирования фильтров, описанной в предыдущем разделе, можно применить метод оптимизации, известный как линейное программирование [19, 20]. Наиболее общая задача линейного программирования формулируется следующим образом:

минимизировать $z = C_1^T X_1 + C_2^T X_2 + d$, (3.9а)

если

p_1: $A_{11}X_1 + A_{12}X_2 \geqslant B_1$, (3.9б)

p_2: $A_{21}X_1 + A_{22}X_2 = B_2$, (3.9в)

n_1: $X_1 \geqslant 0$, (3.9г)

n_2: X_2 свободно, (3.9д)

где p_1 — число ограничений в виде неравенства, p_2 — число ограничений в виде равенства, n_1 — число нормальных случайных переменных, n_2 — число свободных (не имеющих ограничений)

переменных. Целевая функция z состоит из постоянной стоимости d и линейной функции переменных. Векторы C_1 и C_2 (C_1^T и C_2^T — транспонированные C_1 и C_2 соответственно) в совокупности называют вектором стоимости, векторы B_1 и B_2 — вектором ограничений, векторы X_1 и X_2 — вектором программы; совокупность матриц A_{11}, A_{12}, A_{21} и A_{22} называют матрицей коэффициентов [20].

Указанная задача линейного программирования называется основной. С каждой основной задачей связана еще одна задача линейного программирования, называемая двойственной, которая может быть сформулирована следующим образом:

$$\text{минимизировать } z' = -B_1^T U_1 - B_2^T U_2 - d, \tag{3.10а}$$

если

$$p_1: U_1 \geqslant 0, \tag{3.10б}$$

$$p_2: U_2 \text{ свободно}, \tag{3.10в}$$

$$n_1: -A_{11}^T U_1 - A_{21}^T U_2 \geqslant -C_1, \tag{3.10г}$$

$$n_2: -A_{12}^T U_1 - A_{22}^T U_2 = -C_2. \tag{3.10д}$$

В двойственной задаче имеется по одной нормальной случайной переменной для каждого ограничения в виде неравенства основной задачи, по одной свободной переменной для каждого ограничения в виде равенства, по одному ограничению в виде неравенства для каждой нормальной случайной переменной и по одному ограничению в виде равенства для каждой свободной переменной [20].

Между основной и двойственной задачами могут существовать самые различные соответствия [20]. Для обсуждаемой проблемы важное значение имеют соответствия двух типов. Во-первых, легко показать, что задача, двойственная по отношению к двойственной задаче, есть основная задача. Во-вторых, можно показать, что если основная (двойственная) задача имеет оптимальное решение, то двойственная (основная) задача также имеет оптимальное решение, причем

$$z_{\text{опт}} + z'_{\text{опт}} = 0.$$

Таким образом, двойственность позволяет дать две формулировки любой задачи линейной оптимизации. Часто более выгодно решать задачу в одной формулировке и менее выгодно — в другой. В этом мы очень скоро убедимся.

Задачи линейного программирования решают с помощью алгоритма, называемого симплексным [19, 20]. Бóльшая часть подробного описания этого алгоритма в данном контексте не представляет интереса. Следует только отметить, что алгоритм

может дать один из трех взаимно исключающих результатов. Первый результат — оптимальное решение (если задача имеет конечное оптимальное решение); второй результат — указание на то, что задача решаема, но оптимальное решение не имеет границ; третий результат — указание на то, что задача нерешаема. Задача имеет решение, если при некотором (не обязательно оптимальном) выборе переменных можно одновременно удовлетворить условиям всех ограничений. Задача не имеет решения, если при любом выборе переменных невозможно одновременно удовлетворить условиям всех ограничений. В процессе приближения к искомому результату в задачу вводятся дополнительные переменные, называемые произвольными. В случае основной задачи добавляют $p_1 + p_2$ переменных; в случае двойственной задачи $n_1 + n_2$ переменных; с учетом обоих случаев полное число добавляемых переменных составляет $n_1 + n_2 + p_1 + p_2$. Нас не интересуют подробности, касающиеся этих переменных. Однако следует отметить, что если задача имеет оптимальное решение, то число ненулевых переменных, содержащихся в оптимальном векторе программы, составляет $p_1 + p_2$ в случае основной задачи и $n_1 + n_2$ в случае двойственной задачи. Такие ненулевые переменные называют базисными, и они играют важную роль в алгоритме (описанном в п. 3.2.4) для проектирования нерекурсивных цифровых фильтров.

Цель задачи проектирования фильтра состоит в выборе коэффициентов импульсной характеристики в выражениях (3.5)—(3.7), обеспечивающих минимизацию δ в уравнениях

$$H(\mu, \nu) \leqslant F(\mu, \nu) + \delta,$$
$$H(\mu, \nu) \geqslant F(\mu, \nu) - \delta,$$

где $F(\mu, \nu)$ — искомая частотная характеристика, а μ и ν изменяются в пределах некоторого подмножества области $-\pi \leqslant \mu \leqslant \pi$, $-\pi \leqslant \nu \leqslant \pi$. Иногда желательно получить наилучшую аппроксимацию на бесконечном числе точек (μ, ν). Чтобы можно было решать такие задачи с помощью линейного программирования, необходимо выделить из бесконечномерного множества некоторое конечномерное подмножество. Процесс выделения конечномерного подмножества называется отбором, а вносимые при этом ошибки называются ошибками отбора. Можно показать [6], что эти ошибки уменьшаются по мере повышения плотности конечномерного подмножества в бесконечномерном множестве. Влияние эффекта отбора при решении задачи проектирования фильтров рассматривается ниже в п. 3.2.5.

Предположим, что отобрано некоторое конечномерное множество точек p_1 и что $H(\mu, \nu)$ выражается с помощью (3.5). Тогда задачу проектирования фильтра можно сформулировать

как основную задачу линейного программирования следующего вида [см. (3.9)]:

$$\text{минимизировать } z = X_1,$$

если

$$2p_1\colon A_{11}X_1 + A_{12}X_2 \geqslant B_1,$$
$$n_1\colon X_1 \geqslant 0,$$
$$n_2\colon X_2 \text{ свободно,}$$

где

$X_1 = \delta$,

$X_2^T = [h(0, 0), h(1, 0), \ldots, h(N, 0), h(-N, 1), \ldots, h(N, 1), \ldots, \ldots, h(-N, N), \ldots, h(N, N)]$,

$B_1^T = [F(\mu_1\ \nu_1), -F(\mu_1, \nu_1), \ldots, F(\mu_{p1}, \nu_{p1}), -F(\mu_{p1}, \nu_{p1})]$.

A_{11} — матрица размера $2p_1 \times 1$, все элементы которой равны единице; A_{12} — матрица, представленная на фиг. 3.2. Поскольку ограничений в виде равенства нет, $p_2 = 0$. При этом $n_1 = 1$, а n_2 равно числу независимых коэффициентов импульсной характеристики. Задача для (3.6) и (3.7) формулируется совершенно аналогично.

Выше задача проектирования фильтра была сформулирована как основная задача линейного программирования, содержащая $2N^2 + 2N + 2$ переменных и $2p_1$ ограничений. В большинстве задач проектирования фильтров число ограничений обычно превышает число переменных. Однако задача линейного программирования обычно решается гораздо более эффективным способом, если число переменных превышает число ограничений. Для достижения такой ситуации можно сформулировать задачу, двойственную относительно рассматриваемой (основной) задачи [см. (3.10)]:

$$\text{минимизировать } z' = -B_1^T U_1,$$

если

$$2p_1\colon U_1 \geqslant 0,$$
$$n_1\colon -A_{11}^T U_1 \geqslant -1,$$
$$n_2\colon -A_{12}^T U_1 = 0.$$

Как было указано выше, оптимальный вектор программы в случае двойственной задачи линейного программирования содержит $n_1 + n_2$ ненулевых элементов. Следовательно, из матрицы ограничений двойственной задачи можно исключить, не влияя на решение, все столбцы, кроме этих $n_1 + n_2$ столбцов. Более того, оказывается, что условия ограничений основной задачи, соответствующие указанным $n_1 + n_2$ столбцам, удовлетворяются в виде равенства (т. е. $|F(\mu, \nu) - H(\mu, \nu)| = \delta$). Поскольку все

$$
\begin{pmatrix}
1 & 2\cos\mu_1 \ldots & 2\cos N\mu_1 & 2\cos(-N\mu_1+\nu_1)\ldots & 2\cos(N\mu_1+\nu_1)\ldots & 2\cos(-N\mu_1+N\nu_1)\ldots & 2\cos(N\mu_1+N\nu_1)\\
-1 & -2\cos\mu_1 \ldots & -2\cos N\mu_1 & -2\cos(-N\mu_1+\nu_1)\ldots & -2\cos(N\mu_1+\nu_1)\ldots & -2\cos(-N\mu_1+N\nu_1)\ldots & -2\cos(N\mu_1+N\nu_1)\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\
1 & 2\cos\mu_{p_1} \ldots & 2\cos N\mu_{p_1} & 2\cos(-N\mu_{p_1}+\nu_{p_1})\ldots & 2\cos(N\mu_{p_1}+\nu_{p_1})\ldots & 2\cos(-N\mu_{p_1}+N\nu_{p_1})\ldots & 2\cos(N\mu_{p_1}+N\nu_{p_1})\\
-1 & -2\cos\mu_{p_1} \ldots & -2\cos N\mu_{p_1} & -2\cos(-N\mu_{p_1}+\nu_{p_1})\ldots & -2\cos(N\mu_{p_1}+\nu_{p_1})\ldots & -2\cos(-N\mu_{p_1}+N\nu_{p_1})\ldots & -2\cos(N\mu_{p_1}+N\nu_{p_1})
\end{pmatrix}
$$

Фиг. 3.2. Матрица ограничений A_{12}.

$n_1 + n_2$ ненулевых элементов вектора программы положительны и поскольку абсолютное значение ошибки в каждой точке с ограничениями равно δ, ограничения в виде равенства для двойственной задачи (при $U_1 = U_1$ опт.) в явном виде выражают тот факт, что нулевая точка находится на выпуклой оболочке множества $\{r(\mu, \nu)\,[g_1(\mu, \nu), \ldots, g_n(\mu, \nu)] : |r(\mu, \nu)| = \|r(\mu, \nu)\|\}$, как формулирует теорема идентификации (п. 3.1.2).

Автор пользовался комплексом программ для линейного программирования фирмы ИБМ, имеющим название «Система математического программирования 360» (регистрационный номер 360 A-CO-14X), причем применялась ЦВМ типа ИБМ 360/67. В этой системе используется переработанный симплексный алгоритм, причем пользователь может дополнять систему своими собственными программами на ФОРТРАНе. Ряд возможностей системы, основанных на использовании структурных особенностей ЦВМ «Система/360», позволяют быстро находить точные решения задач линейного программирования.

3.2. Алгоритмы

3.2.1. Функции окна

Как отмечалось в п. 3.1.2, многие типы идеальных двумерных фильтров обладают импульсной характеристикой бесконечной протяженности. Это означает, что такие фильтры невозможно реализовать, строго придерживаясь нерекурсивных методов. Один из способов получения нерекурсивной аппроксимации требуемой частотной характеристики состоит в усечении импульсной характеристики, чтобы последняя имела желательную протяженность. Такую процедуру можно рассматривать как операцию умножения импульсной характеристики на функцию окна, имеющую единичное значение в пределах некоторой ограниченной области и нулевое значение вне этой области. Вообще говоря, более качественная аппроксимация достигается в том случае, когда функция окна принимает не только единичное значение в пределах ненулевой области.

Проблему отыскания «хороших» двумерных функций окна изучал Хуанг [21], который доказал следующее положение. Если симметричное одномерное окно $w(x)$ (где x — непрерывная переменная) и двумерное окно $w_2(x, y)$ связаны соотношением

$$w_2(x, y) = w\left(\sqrt{x^2 + y^2}\right),$$

то их преобразования Фурье $W(\mu)$ и $W_2(\mu, \nu)$ удовлетворяют условию

$$\frac{1}{2\pi} W_2(\mu, \nu) \otimes H_2(\mu, \nu) = W(\mu) \otimes H(\mu),$$

где

$$H(\mu)=\begin{cases}1 \text{ при } \mu\geqslant 0,\\ 0 \text{ при } \mu<0,\end{cases}$$

$$H_2(\mu, \nu)=\begin{cases}1 \text{ при } \mu\geqslant 0 \text{ и всех } \nu,\\ 0 \text{ при } \mu<0 \text{ и всех } \nu,\end{cases}$$

а знак $\otimes$ обозначает операцию свертки.

Поскольку полученный результат относится к функциям непрерывной переменной, а не к двум дискретным переменным, его нельзя применить к проектированию двумерных цифровых фильтров, не делая никаких оговорок. Во-первых, следует рассмотреть особенности формирования непрерывной двумерной функции окна из непрерывной одномерной функции окна, согласно выражению $w_2(x, y)=w\left(\sqrt{x^2+y^2}\right)$. Очевидно, что это выражение соответствует вращению одномерной функции окна в xy-плоскости относительно начала координат. Однако это не приводит к простому вращению частотной характеристики, соответствующей рассматриваемому одномерному окну, в $\mu\nu$-плоскости относительно начала координат. Двумерная частотная характеристика, обладающая круговой симметрией, определяется выражением

$$W(\mu, \nu)=W\left(\sqrt{\mu^2+\nu^2}\right)=W(\rho)=2\pi\int_0^\infty w(r)\, r\, J_0(\rho r)\, dr,$$

где $r=\sqrt{x^2+y^2}$, а J_0 — функция Бесселя первого рода нулевого порядка. Такое особое преобразование Фурье для функций с круговой симметрией называется преобразованием Ганкеля [22]. В качестве примера рассмотрим одномерное прямоугольное окно

$$w(t)=\begin{cases}1 \text{ при } |t|\leqslant a,\\ 0 \text{ при других значениях,}\end{cases}$$

частотная характеристика которого имеет вид

$$W(\mu)=\frac{2\sin a\mu}{\mu}.$$

Вращение прямоугольного окна в xy-плоскости относительно начала координат дает функцию

$$w_2(x, y)=\begin{cases}1 \text{ при } \sqrt{x^2+y^2}\leqslant a,\\ 0 \text{ при других значениях,}\end{cases}$$

частотная характеристика которой имеет вид

$$W_2(\mu, \nu) = \frac{2\pi a J_1(a\sqrt{\mu^2+\nu^2})}{\sqrt{\mu^2+\nu^2}}.$$

Эта характеристика подобна частотной характеристике, получаемой путем вращения одномерной частотной характеристики в $\mu\nu$-плоскости относительно начала координат, но, конечно, не совпадает с ней.

Во-вторых, следует рассмотреть эффект дискретизации двумерной функции окна. Дискретизацию можно рассматривать как операцию умножения окна на массив импульсов, что эквивалентно операции свертки частотной характеристики окна с другим массивом импульсов. Поскольку частотная характеристика функции окна не может размещаться в строго ограниченной полосе частот, неизбежно возникают ложные частоты. Этот эффект вынуждает дискретизировать функцию окна с относительно высокой плотностью отсчетов. Таким образом, рассматриваемый метод непригоден для формирования дискретных функций окна с малым числом отсчетов.

Остается рассмотреть то обстоятельство, что частотная характеристика свертывается не со ступенчатой функцией $H_2(\mu, \nu)$, определенной выше, а с фактической частотной характеристикой идеального фильтра. Частотная характеристика «хорошей» функции окна аппроксимирует импульс, т. е. она имеет узкий центральный лепесток и боковые лепестки, уровень которых существенно ниже уровня центрального лепестка. Рассмотрим следующий пример. Требуется применить описанный метод для проектирования фильтра нижних частот с идеальной частотной характеристикой вида

$$H(\mu, \nu) = \begin{cases} 1 \text{ при } \sqrt{\mu^2+\nu^2} \leqslant R, \\ 0 \text{ при других значениях.} \end{cases}$$

В этом случае ширина центрального лепестка частотной характеристики функции окна должна быть много меньше R. Если это условие соблюдается, то циклическая свертка частотной характеристики функции окна с $H(\mu, \nu)$ даст приблизительно тот же результат, что и линейная свертка, согласно соотношению, полученному Хуангом. С такими оговорками справедливо следующее: если $w(x)$ — хорошая одномерная функция окна, то $w_2(x, y) = w(\sqrt{x^2+y^2})$ есть хорошая двумерная функция окна с круговой симметрией.

Рассмотренный метод можно применять для проектирования фильтров в том случае, если известны импульсные характеристики идеального фильтра и функции окна. Однако в большинстве случаев идеальные фильтры задаются их частотными, а не

импульсными характеристиками. Поэтому импульсную характеристику нужно определить с помощью выражения

$$h(n, m) = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} H(\mu, \nu)\, e^{j(\mu n + \nu m)}\, d\mu\, d\nu, \qquad (3.11)$$

где $H(\mu, \nu)$ — идеальная частотная характеристика. Импульсные характеристики идеальных фильтров нескольких типов (фильтра нижних частот, полосового фильтра, фильтра верхних частот и дифференциатора) приведены в п. 3.1.2.

Простейшим двумерным окном является прямоугольное окно с импульсной характеристикой вида

$$w(n, m) = \begin{cases} 1 \text{ при } -N \leqslant n \leqslant N \text{ и } -M \leqslant m \leqslant M, \\ 0 \text{ при других значениях.} \end{cases}$$

Эта функция окна имеет частотную характеристику

$$W(\mu, \nu) = \frac{\sin(2N+1)\frac{\mu}{2}}{\sin\frac{\mu}{2}} \frac{\sin(2M+1)\frac{\nu}{2}}{\sin\frac{\nu}{2}}.$$

Сходство с одномерным случаем очевидно. Как и одномерное прямоугольное окно, данное окно не является очень хорошим. Лучшие двумерные окна можно получить на базе известных одномерных непрерывных функций окна. Одним из простейших таких одномерных окон является окно Хэмминга вида

$$w(t) = \begin{cases} 0{,}54 + 0{,}46\cos(\pi t/\tau) \text{ при } t < \tau, \\ 0 \text{ при других значениях.} \end{cases}$$

Окна Кайзера [3] относятся к более сложным функциям окна, задаваемым выражением

$$w(t) = \begin{cases} \dfrac{I_0\left[\alpha\sqrt{(\tau^2 - t^2)}\right]}{I_0(\alpha t)} \text{ при } |t| < \tau, \\ 0 \text{ при других значениях,} \end{cases}$$

где I_0 — модифицированная функция Бесселя первого рода нулевого порядка, α — константа. Изменяя произведение $\alpha\tau$, можно управлять шириной центрального лепестка частотной характеристики и уровнем боковых лепестков (относительно уровня центрального лепестка).

Легко показать, что перемножение импульсной характеристики идеального фильтра и функции окна соответствует операции циклической свертки их частотных характеристик. Пусть $w(n, m)$

и $f(n, m)$ — импульсные характеристики окна и идеального фильтра соответственно. Тогда импульсная характеристика нерекурсивного фильтра определяется как

$$g(n, m) = w(n, m) \cdot f(n, m).$$

При этом частотная характеристика нерекурсивного фильтра имеет вид

$$G(\mu, \nu) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \left[\frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} W(\alpha, \beta) e^{j(\alpha n + \beta m)} d\alpha d\beta \right] \times$$

$$\times f(n, m) e^{-j(\mu n + \nu m)},$$

где выражение в квадратных скобках заменяет $w(n, m)$. Перестановка операций суммирования и интегрирования и объединение двух экспоненциальных членов дает желаемый результат:

$$G(\mu, \nu) = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} W(\alpha, \beta) F(\mu - \alpha, \nu - \beta) d\alpha d\beta.$$

Известно, что метод функций окна не позволяет проектировать оптимальные фильтры. Более того, этот метод имеет и другие недостатки. Прежде всего бывает трудно, а иногда и просто невозможно найти импульсную характеристику идеального фильтра в замкнутой форме, используя (3.11). Идеальные фильтры, частотные характеристики которых представлены в п. 3.1.2, легко описать аналитически. В некоторой области, которую можно без труда определить, частотная характеристика принимает ненулевое значение; в пределах этой области частотная характеристика легко поддается описанию. Очень легко определить идеальные фильтры, для которых интегрирование в (3.11) чрезвычайно затруднительно. С другой стороны, поскольку большинство двумерных функций окна получают из одномерных функций окна, искомую импульсную характеристику обычно удается без труда выразить в замкнутой форме.

Однако трудности не всегда остаются позади, даже если удается выразить импульсную характеристику в замкнутой форме. Приведенные здесь примеры свидетельствуют о том, что получение отсчетов функции окна может оказаться связанным с вычислением сложных функций, таких, как функции Бесселя и модифицированные функции Бесселя. Конечно, такие вычисления не являются невыполнимыми, однако в случае окон большой протяженности они занимают много времени.

Вторая проблема метода функций окна заключается в том, что на расположение границ полосы пропускания результирующего фильтра весьма трудно влиять. Как уже упоминалось, час-

тотная характеристика нерекурсивного фильтра получается в результате (циклической) свертки частотной характеристики идеального фильтра с частотной характеристикой функции окна. Идеальные фильтры обычно обладают очень крутыми склонами частотной характеристики, однако операция свертки превращает эти склоны в переходные полосы. Таким образом, выбор границ полосы пропускания идеального фильтра следует производить с учетом такого «сглаживания». Кроме того, сглаживающий эффект свертки существенно затрудняет проектирование указанным способом узкополосных или широкополосных нерекурсивных фильтров.

Существует еще одна проблема, упомянутая выше. Дело в том, что фактическая частотная характеристика дискретной функции окна представляет собой модифицированную частотную характеристику соответствующей непрерывной двумерной функции окна, в которой присутствуют ложные частоты. Учет этого обстоятельства обычно сводится к использованию функций окна большей протяженности.

Рабинер [23] применил описанный метод для проектирования нерекурсивных фильтров. Он использовал фильтр нижних частот с круговой симметрией в качестве идеального фильтра и две функции окна: окно с круговой симметрией и резкой границей («нуль — единица») и окно Кайзера. Как и ожидалось, частотная характеристика фильтра, спроектированного с использованием окна Кайзера, имела существенно меньшие пульсации, чем частотная характеристика другого нерекурсивного фильтра.

3.2.2. Частотная дискретизация

Нерекурсивный цифровой фильтр полностью характеризуется либо его импульсной характеристикой, либо эквивалентным числом равномерно расположенных отсчетов его частотной характеристики. Такие отсчеты определяют с помощью дискретного преобразования Фурье. В основе частотной дискретизации как методе проектирования нерекурсивных фильтров лежит следующий факт: (непрерывную) частотную характеристику нерекурсивного фильтра можно записать в виде линейной функции указанных частотных отсчетов. Следовательно, можно оптимизировать частотную характеристику в том или ином отношении, применив один из методов линейной оптимизации, например линейное программирование. Обратное ДПФ позволяет затем рассчитать импульсную характеристику фильтра. Как правило, в качестве переменных выбирают лишь небольшое число частотных отсчетов; остальным отсчетам придаются значения, которые имеет идеальный фильтр на соответствующих частотах. Об этом ограничении будет сказано подробнее несколько ниже.

Для того чтобы показать, каким образом частотная характеристика нерекурсивного фильтра записывается в виде линейной функции ее частотных отсчетов, полезно ознакомиться с вариантом ДПФ, несколько отличающимся от варианта, приведенного в п. 3.1.1. Предположим, что импульсная характеристика фильтра принимает нулевое значение вне области $-N \leqslant n \leqslant N$ и $-N \leqslant m \leqslant N$. При этом z-преобразование импульсной характеристики определяется выражением

$$H(z_1, z_2) = \sum_{n=-N}^{N} \sum_{m=-N}^{N} h(n, m)\, z_1^{-n} z_2^{-m}. \tag{3.12}$$

ДПФ, соответствующее равномерно расположенным отсчетам z-преобразования, определяется как

$$H(k, l) = \sum_{n=-N}^{N} \sum_{m=-N}^{N} h(n, m) \exp\left[-j \frac{2\pi}{Q} (nk + ml)\right],$$

где $Q = 2N + 1$, $-N \leqslant k \leqslant N$ и $-N \leqslant l \leqslant N$. Легко проверить, что в этом случае обратное ДПФ определяется выражением

$$h(n, m) = \frac{1}{Q^2} \sum_{k=-N}^{N} \sum_{l=-N}^{N} H(k, l) \exp\left[j \frac{2\pi}{Q} (nk + ml)\right]. \tag{3.13}$$

Выражения (3.13) и (3.12) в совокупности дают следующий результат:

$$H(z_1, z_2) = \frac{z_1^N z_2^N \left(1 - z_1^{-Q}\right)\left(1 - z_2^{-Q}\right)}{Q^2} \times$$

$$\times \sum_{k=-N}^{N} \sum_{l=-N}^{N} \frac{H(k, l) \exp\left[j \frac{2\pi}{Q} N(k + l)\right]}{\left[1 - z_1^{-1} \exp\left(-j \frac{2\pi}{Q} k\right)\right]\left[1 - z_2^{-1} \exp\left(-j \frac{2\pi}{Q} l\right)\right]}.$$

Частотная характеристика фильтра, полученного при условии $z_1 = \exp(j\mu)$ и $z_2 = \exp(j\nu)$, имеет вид

$$H(\mu, \nu) = \frac{e^{jN(\mu+\nu)}\left(1 - e^{-j\mu Q}\right)\left(1 - e^{-j\nu Q}\right)}{Q^2} \times$$

$$\times \sum_{k=-N}^{N} \sum_{l=-N}^{N} \frac{H(k, l) \exp\left[j \frac{2\pi}{Q} (k + l)\right]}{\left\{1 - \exp\left[-j\left(\mu + \frac{2\pi}{Q} k\right)\right]\right\}\left\{1 - \exp\left[-j\left(\nu + \frac{2\pi}{Q} l\right)\right]\right\}}. \tag{3.14}$$

Чтобы преобразовать (3.14) к виду, пригодному для применения линейного программирования, необходимо ввести два огра-

ничения, а именно следует считать, что импульсная и частотные характеристики являются чисто действительными. При введении этих двух ограничений отсчеты ДПФ удовлетворяют условию

$$H(k, l) = H(-k, -l).$$

Используя это соотношение между частотными отсчетами, можно записать (3.14) в виде

$$H(\mu, \nu) = \frac{1}{Q^2}\left\{ H(0, 0)\,\phi_1(\mu)\,\phi_1(\nu) + \right.$$

$$+ \sum_{k=1}^{N} H(k, 0)\,\phi_1(\nu)\,[\phi_2(\mu, k) + \phi_2(\mu, -k)] +$$

$$\left. + \sum_{k=-N}^{N} \sum_{l=1}^{N} H(k, l)\,[\phi_2(\mu, k)\,\phi_2(\nu, l) + \phi_2(\mu, -k)\,\phi_2(\nu, -l)] \right\}, \quad (3.15)$$

где

$$\phi_1(x) = \frac{\sin Q\frac{x}{2}}{\sin\frac{x}{2}},$$

$$\phi_2(x, y) = \frac{\sin\left[Q\left(\frac{x}{2} + \frac{\pi y}{Q}\right)\right]}{\sin\left(\frac{x}{2} + \frac{\pi y}{Q}\right)}.$$

На первый взгляд (3.15) не имеет преимуществ по сравнению с (3.14). Однако важно отметить, что теперь все коэффициенты частотных отсчетов являются действительными, и это позволяет применить линейное программирование.

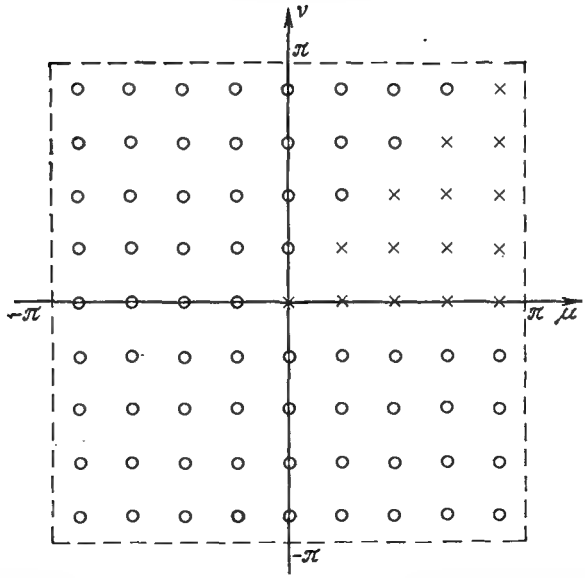
Число независимых частотных отсчетов можно уменьшить, если на частотную характеристику фильтра наложить ограничения по симметрии. В частности, если искомая частотная характеристика обладает круговой симметрией, на частотные отсчеты может быть наложено следующее ограничение:

$$H(k, l) = H(|k|, |l|)$$

и

$$H(k, l) = H(l, k).$$

Этот случай иллюстрирует фиг. 3.3, где показан массив 9×9 частотных отсчетов. При использовании метода частотной дискретизации число изменяемых частотных отсчетов обычно ограничивают в еще большей степени, приписывая большинству отсчетов постоянное значение. Если речь идет о фильтре нижних частот, то отсчеты в пределах полосы пропускания устанавливают



Фиг. 3.3. Ограничения по симметрии в частотной области; × — независимые частотные отсчеты; О — частотные отсчеты, обусловленные ограничениями по симметрии.

равными единице, а отсчеты в полосе задерживания выбирают равными нулю; только в переходной полосе используются изменяемые отсчеты. На фиг. 3.4 в качестве примера представлена частотная характеристика фильтра нижних частот размерности 9×9. Если в переходной полосе имеется M частотных отсчетов, то (3.15) можно переписать в виде

$$H(\mu, \nu) = H_F(\mu, \nu) + \sum_{k=1}^{M} T_k H_k(\mu, \nu), \tag{3.16}$$

где $H_F(\mu, \nu)$ учитывает вклад частотных отсчетов с фиксированным значением, а T_k обозначает отсчеты переходной полосы; сумма учитывает вклад отсчетов переходной полосы в формировании частотной характеристики. В общем случае частотные характеристики, вычисленные указанным методом, не являются наилучшими аппроксимациями искомой частотной характеристики, поскольку могут меняться только частотные отсчеты переходной полосы.

Выбор оптимальных значений частотных отсчетов в переходной полосе и минимизация пульсаций в полосах пропускания и задерживания могут быть осуществлены с помощью линейного

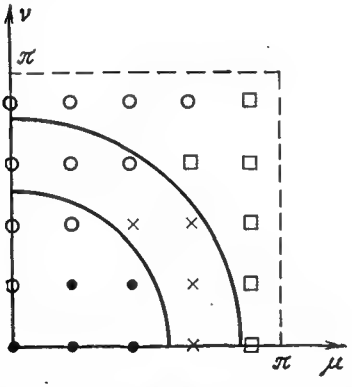
программирования. Для этого требуется, чтобы равенства ограничений имели вид

$$H(\mu, \nu) \leqslant F(\mu, \nu) + \delta,$$
$$H(\mu, \nu) \geqslant F(\mu, \nu) - \delta,$$

где $H(\mu, \nu)$ определяется согласно (3.16), а $F(\mu, \nu)$ — искомая частотная характеристика. Для каждой точки с ограничениями, находящейся в полосе пропускания или полосе задерживания, следует записать два ограничения в виде равенства. Для точек в переходной полосе ограничения в виде равенства не записываются. Точки с ограничениями в частотной плоскости должны располагаться настолько плотно, чтобы в этих точках обеспечивалось хорошее приближение результирующей непрерывной частотной характеристики к искомой частотной характеристике. Однако плотность множества таких точек не должна быть излишне высокой, иначе вычислительные трудности могут сделать задачу нерешаемой. Ху [24, 25] установил, что если массив частотных отсчетов имеет размерность $M \times M$, то достаточно иметь $8M \times 8M$ точек с ограничениями. Ограничения по симметрии, о которых говорилось выше, позволяют пользоваться только одной восьмой указанного числа точек с ограничениями.

После определения оптимальных значений частотных отсчетов в переходной полосе можно вычислить импульсную характеристику фильтра, используя обратное ДПФ, согласно (3.13).

Ху [25] использовал описанный метод для проектирования разнообразных фильтров нижних частот с круговой симметрией.



Фиг. 3.4. Размещение частотных отсчетов в случае фильтра нижних частот; × — изменяемые частотные отсчеты, ● — частотные отсчеты, принимаемые за единицу; □ — частотные отсчеты, принимаемые за нуль; ○ — частотные отсчеты, обусловленные ограничениями по симметрии.

Каждый такой фильтр имел массив частотных отсчетов размерности 25×25, однако принятое ограничение по симметрии позволило считать потенциально независимыми переменными только 91 отсчет. Большинству из этих 91 отсчета было придано значение единица или нуль; число переменных частотных отсчетов в переходной полосе изменялось от 2 до 16. Ширина переходной полосы изменялась в пределах от $0{,}08\pi$ до $0{,}24\pi$, а граница полосы пропускания изменялась от нуля до $0{,}88\pi$. Пульсации в полосе задерживания минимизировались без наложения ограничений на пульсации в полосе пропускания. Затухание в полосе задерживания изменялось от 10 дБ (узкие переходные полосы) до приблизительно 70 дБ (широкие переходные полосы). Число точек с ограничениями изменялось приблизительно от 1500 до 5000 и зависело от положения и ширины переходной полосы. Вычисления занимали от 20 до 30 мин. Работа была выполнена на ЦВМ GE 635 с использованием комплекса линейного программирования, имеющего название «Подпрограмма АПММ фирмы ИБМ для научных задач» [26].

3.2.3. Прямое применение линейного программирования

Как уже указывалось, если импульсную и частотную характеристики нерекурсивного цифрового фильтра считать чисто действительными, то (непрерывную) частотную характеристику можно записать в виде линейной функции коэффициентов импульсной характеристики или частотных отсчетов ДПФ. В первом случае частотная характеристика определяется согласно (3.5). Соответствие между частотной характеристикой и частотными отсчетами ДПФ дает (3.15). Пользуясь этими соотношениями, нетрудно записать равенства ограничений следующего вида:

$$H(\mu, \nu) \leqslant F(\mu, \nu) + \delta,$$
$$H(\mu, \nu) \geqslant F(\mu, \nu) - \delta,$$

где $F(\mu, \nu)$ — искомая частотная характеристика. Линейное программирование позволяет минимизировать пульсации δ с учетом матрицы ограничений, которая определяется путем составления пары уравнений ограничений для каждой точки с ограничениями. Точки с ограничениями должны быть расположены достаточно плотно, чтобы можно было считать, что значения частотной характеристики между точками с ограничениями близки к ее значениям в этих точках, но не настолько плотно, чтобы задача была нерешаемой вследствие вычислительных трудностей.

Уравнения ограничений можно также записать в более общем виде:

$$H(\mu, \nu) \leqslant F(\mu, \nu) + W(\mu, \nu)\delta,$$
$$H(\mu, \nu) \geqslant F(\mu, \nu) - W(\mu, \nu)\delta,$$

где $W(\mu, \nu)$ — действительная весовая функция. Последняя позволяет влиять на функцию ошибки полученной аппроксимации в желательном направлении. Если речь идет о фильтрах (таких, как фильтры нижних и верхних частот и полосовые фильтры) с резко ограниченными полосами пропускания и задерживания, то можно использовать весовую функцию для получения различных пульсаций в этих полосах. Соответствующая весовая функция определяется в виде

$$W(\mu, \nu) = \begin{cases} 1 \text{ для точек с ограничениями в полосе пропускания,} \\ \alpha \text{ для точек с ограничениями в полосе задерживания.} \end{cases}$$

Величина результирующих пульсаций в полосе задерживания в α раз (самое большее) превышает величину результирующих пульсаций в полосе пропускания.

Пусть импульсная характеристика фильтра имеет нулевое значение вне области $-N \leqslant n \leqslant N$, $-N \leqslant m \leqslant N$, а импульсная и частотная характеристики являются действительными. При таком условии число независимых коэффициентов импульсной характеристики или частотных отсчетов ДПФ равно $2N^2 + 2N + 1$, причем точки с ограничениями должны заполнять область $-\pi \leqslant \mu \leqslant \pi$, $0 \leqslant \nu \leqslant \pi$. Вообще говоря, для получения наилучшей аппроксимации искомой частотной характеристики необходимо использовать все эти переменные. Однако с увеличением N число переменных быстро растет, а время, которое требуется для получения решения с помощью линейного программирования, возрастает еще быстрее. Следовательно, целесообразно наложить рассмотренные выше ограничения по симметрии, что позволит уменьшить число переменных и требуемое число точек с ограничениями. Если принять, что $h(n, m) = h(|n|, |m|)$, то число переменных уменьшается до $N^2 + 2N + 1$, а область, которую должны заполнить точки с ограничениями, сокращается вдвое. Дополнительное допущение $h(n, m) = h(m, n)$ приводит к уменьшению числа переменных до $(N^2 + 3N + 2)/2$ и к новому сокращению вдвое площади, заполняемой точками с ограничениями. Наложение всех этих ограничений по симметрии обеспечивает резкое сокращение времени вычислений. Дополнительные сведения по этому вопросу приводятся ниже.

Если принять лишь одно допущение о действительности импульсной и частотной характеристик фильтра, то точки с ограничениями должны плотно заполнять область $-\pi \leqslant \mu \leqslant \pi$, $0 \leqslant \nu \leqslant \pi$. Практически при вычислении «непрерывной» частотной характеристики фильтра используется 256×256-точечное БПФ, поэтому фактически заполняется область $-\pi \leqslant \mu \leqslant (127/128)\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. Точки с ограничениями расположены в узлах прямоугольной решетки с шагом π/KN, причем принимается, что импульсные характеристики простираются от $-N$ до N в каждом из двух направлений, K — целочисленная константа (обычно выбираемая в пределах от 5 до 15). Рассмотрим следующий пример. Требуется спроектировать фильтр нижних частот с круговыми границами частотной характеристики, причем радиус границы полосы пропускания составляет R_1, а радиус границы полосы задерживания R_2. В качестве точек с ограничениями фактически используются только те точки решетки, которые одновременно удовлетворяют двум условиям вида

$$\sqrt{\mu^2+\nu^2} < R_1 - \frac{\pi}{KN}, \tag{3.17а}$$

$$\sqrt{\mu^2+\nu^2} > R_2 + \frac{\pi}{KN}. \tag{3.17б}$$

Кроме того, имеются точки с ограничениями, расположенные непосредственно на каждой из двух границ. Полное число точек с ограничениями, принадлежащих границам полос, равно целочисленной части результата

$$Q_1 = R_1(KN+1)+1$$

для границы полосы пропускания и

$$Q_2 = R_2(KN+1)+1$$

для границы полосы задерживания. Эти точки равномерно размещены на границах с угловым шагом

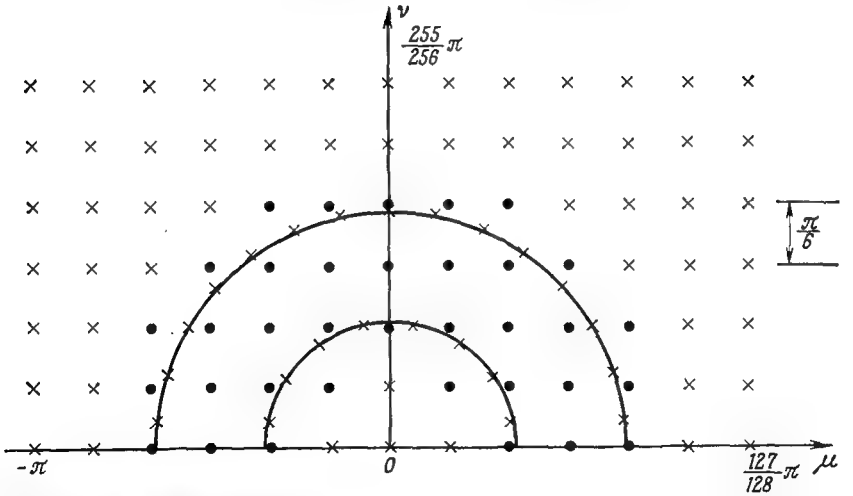
$$\Delta\theta_1 = \frac{\pi}{Q_1} \quad \text{и} \quad \Delta\theta_2 = \frac{\pi}{Q_2},$$

причем начальный угол равен

$$\theta_1 = 0{,}5\frac{\pi}{Q_1} \quad \text{и} \quad \theta_2 = 0{,}5\frac{\pi}{Q_2}.$$

На фиг. 3.5 показан пример размещения точек с ограничениями для $N = 2$, $K = 3$, $R_1 = 1{,}1$ и $R_2 = 2{,}0$.

Если принимается, что импульсная и частотная характеристики являются действительными и, кроме того, удовлетворяют



Фиг. 3.5. Размещение точек с ограничениями (случай 1): × — точки с ограничениями; ● — неиспользуемые точки решетки.

условиям $h(n, m) = h(|n|, |m|)$ и $h(n, m) = h(m, n)$, то точками с ограничениями заполняется вдвое меньшая область $0 \leqslant \mu \leqslant \pi$, $0 \leqslant \nu \leqslant \pi$, Практически частотная характеристика вычисляется с использованием 256×256 точек БПФ, распределенных по области $0 \leqslant \mu \leqslant (255/256)\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. Как и в предыдущем случае, точки с ограничениями размещены в узлах прямоугольной решетки с шагом π/KN. Более того, если речь идет о фильтре нижних частот, то в качестве точек с ограничениями фактически используются только те точки, которые одновременно удовлетворяют обоим условиям (3.17а) и (3.17б). Как и прежде, также имеются точки с ограничениями непосредственно на границах полос. В данном случае полное число точек с ограничениями на границах равно целочисленной части результата

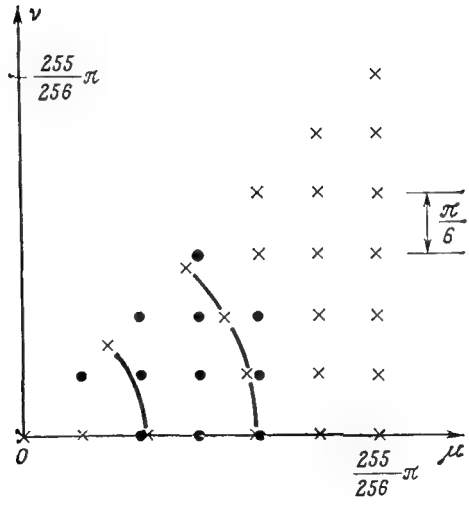
$$Q_1 = \frac{KNR_1}{4} + 1$$

для границы полосы пропускания и

$$Q_2 = \frac{KNR_2}{4} + 1$$

для границы полосы задерживания. Если вычисленное по этим формулам Q_1 или Q_2 оказывается меньше двух, ему придают значение два. Точки равномерно размещены на границах полос с угловым шагом (начальный угол равен нулю)

$$\Delta\theta_1 = \frac{\pi}{4(Q_1 - 1)}, \quad \Delta\theta_2 = \frac{\pi}{4(Q_2 - 1)}.$$



Фиг. 3.6. Размещение точек с ограничениями (случай 2): × — точки с ограничениями; ● — неспользуемые точки решетки.

На фиг. 3.6 показан пример размещения точек с ограничениями для рассматриваемого случая при $N = 2$, $K = 3$, $R_1 = 1{,}1$ и $R_2 = 2{,}0$.

Во всех фильтрах, спроектированных автором указанным методом, частотная характеристика выражалась в виде линейной функции коэффициентов импульсной характеристики. Однако Ху [25] приводит примеры спроектированных им фильтров, для которых частотная характеристика выражается в виде линейной функции частотных отсчетов ДПФ [см. (3.15)]. При этом использовалось то же плотное множество точек с ограничениями, которое было использовано при проектировании ряда фильтров методом частотной дискретизации. Если массив частотных отсчетов имеет размерность $M \times M$, то точки с ограничениями следует выбирать из $8M \times 8M$ точек, составляющих решетку. В качестве точек с ограничениями используют только точки, расположенные на половине одного квадранта, а точки решетки в переходной полосе не учитывают.

Собственно оптимизация выполняется после выбора числа и положения точек с ограничениями. Конкретная реализация этой операции, однако, существенно зависит от особеностей используемого комплекса линейного программирования, поэтому никаких подробных сведений здесь не приводится. Автор пользовался комплексом программ линейного программирования, который является одной из программных систем фирмы ИБМ, нося-

щей название «Система математического программирования/360» (регистрационный номер 360 А-СО-14Х). Использовалась ЦВМ типа ИБМ 360/67.

3.2.4. Новый алгоритм

Существует несколько методов проектирования оптимальных одномерных нерекурсивных цифровых фильтров, которые являются итеративными по своей природе и позволяют найти наилучшую аппроксимацию требуемой частотной характеристики на интервале $[0, 2\pi]$, пользуясь при каждой итерации лишь очень небольшим числом точек. Таким свойством обладают второй алгоритм Ремеза [6], алгоритм Хофстеттера [8] и метод Паркса и Макклеллана [9]. К сожалению, ни один из этих методов нельзя непосредственно распространить на двумерный случай. Во всех трех методах в явном или неявном виде выдвигается требование о том, чтобы функции, используемые в аппроксимации, удовлетворяли условию Хаара; в случае проектирования двумерных фильтров это требование невыполнимо (см. п. 3.1.2). Следует также учесть, что алгоритм Хофстеттера и метод Паркса и Макклеллана основаны на интерполяционной формуле Лагранжа, которую нельзя обобщить на двумерный случай общего вида [27—29]. Однако если бы удалось разработать метод решения задачи двумерной аппроксимации, сходный с перечисленными методами, то это позволило бы значительно уменьшить число ограничений в виде равенства при использовании линейного программирования за счет увеличения числа промежуточных итераций, ведущих к нахождению действительно наилучшей аппроксимации требуемой частотной характеристики.

Поиски такого итерационного метода для многомерного случая основаны на факте существования соответствующих одномерных методов, а также на следующем утверждении. Если обобщенный полином

$$P(x)=\sum_{i=1}^{n} c_i g_i(x)$$

(обозначения приведены в п. 3.1.2)есть наилучшая аппроксимация функции F на множестве X, то существует конечномерное подмножество X_0, включающее самое большее $n+1$ точек, таких, что P является наилучшей аппроксимацией F на X_0 [6]. Почти полное описание итерационного алгоритма для многомерного случая содержится в конце п. 3.1.2. Здесь представлены недостающие подробности.

При выполнении каждой итерации данного алгоритма находится наилучшая аппроксимация требуемой частотной характеристики (с использованием линейного программирования и

представления частотной характеристики в виде функции коэффициентов импульсной характеристики) на «сокращенном» множестве точек с ограничениями. В общем случае максимальное значение ошибки на сокращенном множестве точек с ограничениями определенно меньше максимального абсолютного значения ошибки по всей частотной области, представляющей интерес, а получаемая аппроксимация не является наилучшей аппроксимацией искомой частотной характеристики. Требуется выбрать новое сокращенное множество точек с ограничениями, такое, чтобы максимальное абсолютное значение ошибки на этом множестве превышало максимальное абсолютное значение ошибки на прежнем сокращенном множестве (подразумевается, что на новом множестве точек с ограничениями найдено оптимальное решение).

Чтобы начать указанную процедуру, необходимо иметь исходное решение по выбору сокращенного множества точек с ограничениями. Размещение исходных точек с ограничениями не является критичным. Метод определения местоположений точек совершенно не отличается от того, который описан в п. 3.2.3 и был использован для нахождения плотного множества точек с ограничениями. Точки с ограничениями располагаются в узлах прямоугольной решетки с шагом π/KN, причем принимается, что в каждом из двух направлений импульсная характеристика простирается от $-N$ до N и что K — целочисленная константа. В данном случае константа K выбирается равной двум; однако в других случаях, когда аппроксимация определяется путем решения только одной задачи линейного программирования, выбираются бóльшие значения этой константы. На границах полос точки с ограничениями размещаются точно так же, как описано выше и проиллюстрировано на фиг. 3.5 и 3.6. Как и прежде, точки решетки в переходной полосе не используются в качестве точек с ограничениями. Если не считать выбора исходных точек с ограничениями, все итерации производятся идентично; ниже приводится описание типичной итерации.

Предполагая, что уже найдено решение на некотором множестве точек с ограничениями, опишем основные операции, которые выполняются в процессе каждой итерации.

1. Сохранение данных решения задачи линейного программирования. Представляют интерес следующие компоненты решения: коэффициенты импульсной характеристики, максимальное абсолютное значение ошибки и переменные, составляющие базис для двойственной задачи (п. 3.1.3). Базисные переменные используются при определении очередного множества точек с ограничениями.

2. Вычисление «непрерывной» частотной характеристики в соответствующей части области частот. Практически частотная

характеристика вычисляется по 256×256 точкам решетки с использованием БПФ.

3. Вычисление значений частотной характеристики вдоль границ полос (предполагается, что такие границы имеются, как в случае фильтра нижних частот). Такое вычисление также производится на плотном множестве точек, однако, к сожалению, применить БПФ при этом нельзя.

4. Нахождение локальных максимумов и минимумов функции ошибки в представляющей интерес частотной области, включая границы полос. Эта операция существенно важна для определения очередного множества точек с ограничениями. Следует также заметить, что она является, безусловно, наименее определенной частью нового алгоритма.

5. Построение различных графиков с выходными данными. Такие графики позволяют следить за ходом реализации алгоритма от итерации к итерации.

6. Получение данных для очередной задачи линейного программирования и решение этой задачи.

Первый этап алгоритма заключается в сохранении соответствующих данных решения последней задачи линейного программирования. Должны быть подготовлены следующие данные: коэффициенты импульсной характеристики, максимальное абсолютное значение ошибки и переменные, составляющие базис для двойственной задачи (п. 3.1.3). Решение, обеспечиваемое комплексом линейного программирования «Система математического программирования/360», содержит как оптимальный вектор программы для задачи линейного программирования, которая была решена (основная или двойственная задача), так и оптимальный вектор программы для задачи, двойственной к задаче линейного программирования, которая была решена. Если задача линейного программирования была сформулирована в основной форме, то компонентами оптимального вектора программы являются коэффициенты импульсной характеристики и максимальное абсолютное значение ошибки. Максимальное абсолютное значение ошибки также является оптимальным значением целевой функции. Базисные переменные для двойственной задачи являются ненулевыми компонентами оптимального вектора программы для задачи, двойственной к основной задаче. Если задача линейного программирования была сформулирована в двойственной форме, ненулевыми компонентами оптимального вектора программы являются базисные переменные для двойственной задачи. Коэффициенты импульсной характеристики и максимальное абсолютное значение ошибки являются отрицательными значениями компонентов оптимального вектора программы для задачи, двойственной к двойственной задаче. Максимальное абсолютное значение ошибки также

является отрицательным оптимальным значением целевой функции.

Коэффициенты импульсной характеристики требуется сохранить для того, чтобы можно было вычислить частотную характеристику. Максимальное абсолютное значение ошибки сохраняется при каждой итерации с той целью, чтобы его можно было сравнить со значением, полученным в результате выполнения предшествующей итерации. Если это значение ошибки уменьшается, выполнение алгоритма прекращается. Теоретически значение ошибки должно увеличиваться при переходе от данной итерации к следующей. Так почти всегда и происходит. Тем не менее из 250 (ориентировочно) выполненных итераций обнаружились три, давшие некоторое уменьшение этого значения. Однако следует указать, что во всех трех случаях алгоритм уже почти сходился, когда наступало указанное уменьшение абсолютного значения ошибки. Поэтому это явление было объяснено наличием накопленной вычислительной погрешности. Базисные переменные для двойственной задачи следует сохранить для того, чтобы соответствующие точки с ограничениями можно было включить в множество точек с ограничениями для следующей итерации. Одно из требований теоретического характера к новому множеству точек с ограничениями заключается в том, что система уравнений

$$\Delta = F(\mu, \nu) - \sum_{i=1}^{n} c_i g_i(\mu, \nu)$$

должна быть ограничивающей на этом новом множестве точек (п. 3.1.2). Эта система уравнений является ограничивающей, если все значения Δ отличаются от нуля и если невозможно одновременно уменьшить абсолютное значение всех Δ при произвольном выборе множества $\{c_1, \ldots, c_n\}$. Можно не сомневаться в том, что все значения Δ являются ненулевыми, поскольку абсолютное значение каждой разности Δ равно максимальному абсолютному значению ошибки. Кроме того, все значения Δ невозможно одновременно уменьшить по абсолютному значению, поскольку решение предыдущей задачи линейного программирования оптимально. Любая система уравнений, содержащая некоторую ограничивающую систему, сама является ограничивающей. Поэтому совершенно очевидно, что указанное выше теоретическое требование к множеству точек с ограничениями для очередной итерации будет удовлетворяться в том случае, когда в это новое множество будут включены точки с ограничениями, соответствующие базисным переменным. После каждой очередной итерации положения точек с ограничениями, соответствующих базисным переменным, сравнивают с положениями тех же точек после предыдущей итерации. Если все точки сохра-

няют свое положение, то имеет место сходимость алгоритма и его выполнение прекращают. Это обычная процедура окончания.

Вторая операция алгоритма состоит в вычислении «непрерывной» частотной характеристики фильтра. С помощью БПФ вычисляются 256×256 отсчетов частотной характеристики. Решение об использовании размерности 256×256 массива является компромиссным. С одной стороны, массив должен быть достаточно плотным, чтобы можно было достаточно точно определить положение локальных максимумов и минимумов функции ошибки (четвертый этап алгоритма). С другой стороны, массив должен достаточно хорошо соответствовать имеющейся емкости памяти ЦВМ. Искомое двумерное БПФ вычисляется с использованием одномерного БПФ, как описывается в п. 3.1.1. Используется упрощенный вариант алгоритма БПФ, разработанного Бреннером [30]. Алгоритм был переработан для выполнения одномерного 256-точечного прямого преобразования с удвоенной точностью. До получения окончательного результата все вычисления БПФ производят в арифметическом устройстве, которое обеспечивает удвоенную точность. Окончательный результат, однако, записывается в память в виде массива размерности 256×256 с обеспечением «однократной» точности. Благодаря этому минимизируется влияние вычислительных погрешностей на точность определения локальных минимумов и максимумов функции ошибки (четвертый этап алгоритма).

Если принимается лишь допущение о действительности импульсной и частотной характеристик, то отсчеты частотной характеристики следует равномерно распределить в пределах области $-\pi \leqslant \mu \leqslant (127/128)\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. В этом случае коэффициенты импульсной характеристики удовлетворяют условию симметрии $h(n, m) = h(-n, -m)$. Это условие позволяет ограничиться вычислением только $N+1$ (одномерных) преобразований столбцов. Остальные N преобразований столбцов можно определить на основе симметрии. 256 отсчетов частотной характеристики в области $0 \leqslant \nu \leqslant (255/256)\pi$ можно определить, считая их первой половиной 512-точечного преобразования. Такое 512-точечное преобразование можно вычислить как преобразование двух 256-точечных последовательностей, если воспользоваться фундаментальным понятием алгоритма БПФ с прореживанием по времени. Поскольку такие последовательности являются действительными, они могут быть объединены, а их преобразования могут быть вычислены одновременно; после завершения вычислений результаты разделяются методом выделения четной и нечетной частей. После вычисления преобразований столбцов необходимо перейти к вычислению преобразований строк. Каждая строка является (в общем случае) комплексной. Однако, поскольку известно, что преобразование каж-

дой строки дает действительный результат, пары строк можно объединить и преобразовать совместно. Результаты разделяются после выполнения вычислений. Таким образом, в общей сложности требуется вычислить $129 + N$ одномерных 256-точечных преобразований.

Если принимается, что импульсная и частотная характеристики являются действительными и, кроме того, удовлетворяют условиям $h(n, m) = h(|n|, |m|)$ и $h(n, m) = h(m, n)$, то отсчеты частотной характеристики равномеро размещают в пределах области $0 \leqslant \mu \leqslant (255/256)\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. В данном случае преобразования столбцов вычисляются точно так же, как и в предыдущем. Путем усложнения программы можно добиться небольшого выигрыша. Для этого нужно учесть то обстоятельство, что теперь каждый столбец является не только действительным, но и четным, а следовательно, преобразование каждого столбца является действительным и четным. Однако, чтобы обеспечить эффективность вычисления преобразований строк, следует воспользоваться тем дополнительным обстоятельством, что каждая строка также является действительной и четной. И в этом случае искомые отсчеты частотной характеристики считаются первой половиной 512-точечного преобразования. Однако существует метод [31] вычисления преобразования двух действительных четных 512-точечных последовательностей с помощью одного 256-точечного преобразования. Следовательно, в данном случае полное число одномерных 256-точечных преобразований, подлежащих вычислению, также составляет $129 + N$.

Третий этап алгоритма — вычисление значений частотной характеристики фильтра на границах полос. Многие типы фильтров (например, фильтры нижних и верхних частот и полосовые фильтры) обладают четкими границами полос частотной характеристики. Однако существуют фильтры, подобные упомянутому выше дифференциатору, которые не имеют границ полос частотной характеристики; во всех таких случаях данный этап алгоритма пропускается. В общем случае граница может иметь произвольную форму. Однако все фильтры (нижних частот), спроектированные с использованием данного алгоритма, имеют круговые границы полос. Поэтому для выполнения соответствующих вычислений невозможно воспользоваться алгоритмом БПФ. В данном случае частотная характеристика, выраженная через коэффициенты импульсной характеристики, должна находиться путем прямых вычислений. Однако существенного снижения эффективности алгоритма при этом не происходит, поскольку приходится вычислять лишь небольшое число отсчетов частотной характеристики.

Если принято лишь одно допущение о действительности импульсной и частотной характеристик, то должны быть вычис-

лены отсчеты частотной характеристики, расположенные в пределах дуги окружности 180°. Пусть круговые границы полос пропускания и задерживания фильтра имеют радиусы R_1 и R_2 соответственно. Полное число вычисляемых отсчетов частотной характеристики тогда составляет

$$Q_1 = 2 \cdot (\text{целочисленная часть } 64R_1) + 1$$

для границы полосы пропускания и

$$Q_2 = 2 \cdot (\text{целочисленная часть } 64R_2) + 1$$

для границы полосы задерживания. Эти отсчеты размещены равномерно по границам с угловым шагом

$$\Delta\theta_1 = \frac{\pi}{(Q_1 - 1)}, \qquad \Delta\theta_2 = \frac{\pi}{(Q_2 - 1)}.$$

Поскольку гарантируется нечетность Q_1 и Q_2, всегда приходится вычислять отсчет при углах 0, 90 и 180°.

Если принимается, что импульсная и частотная характеристики являются действительными и, кроме того, удовлетворяют условиям $h(n, m) = h(|n|, |m|)$ и $h(n, m) = h(m, n)$, то частотная характеристика вычисляется только в пределах дуги окружности 45°. В этом случае полное число вычисляемых отсчетов частотной характеристики составляет

$$Q_1 = 2 \cdot (\text{целочисленная часть } 16R_1) + 1$$

для границы полосы пропускания и

$$Q_2 = 2 \cdot (\text{целочисленная часть } 16R_2) + 1$$

для границы полосы задерживания. Отсчеты размещены с угловым шагом

$$\Delta\theta_1 = \frac{\pi}{4(Q_1 - 1)}, \qquad \Delta\theta_2 = \frac{\pi}{4(Q_2 - 1)}.$$

Первый отсчет всегда соответствует углу 0°, последний — углу 45°.

Четвертый этап алгоритма заключается в определении локальных максимумов и минимумов функции ошибки. Точка, в которой текущая функция ошибки имеет локальный максимум или локальный минимум, включается в множество точек с ограничениями для следующей итерации, если абсолютное значение ошибки в этой точке превышает максимальное абсолютное значение ошибки на текущем множестве точек с ограничениями или совпадет с ним. При этом гарантируется удовлетворение двух условий

$$\max_{X_{k+1}} |P_k - F| = \|E_k\|,$$

$$\min_{X_{k+1}} |P_k - F| \geqslant \delta_k,$$

сформулированных при теоретическом обосновании алгоритма (п. 3.1.2). Полное множество точек с ограничениями для следующей итерации составляется из точек этих локальных максимумов и минимумов и упомянутых выше точек с ограничениями, соответствующих базисным переменным двойственной задачи. На первый взгляд может показаться, что нахождение локальных максимумов и минимумов является достаточно простой операцией. Оказывается, однако, что она является наименее определенной частью алгоритма.

Во-первых, было установлено, что при прямом сравнении абсолютного значения ошибки в точке каждого локального максимума или минимума с δ_k (т. е. с максимальным абсолютным значением ошибки на текущем множестве точек с ограничениями) существует опасность ошибочного исключения некоторых точек. Речь идет о точках, в которых вычисленное абсолютное значение ошибки лишь немного меньше δ_k. Чтобы исправить такое положение, абсолютное значение ошибки в каждой точке сравнивается не с δ_k, а с $\delta_T = 0{,}999\ \delta_k$. Это решение оказалось вполне удовлетворительным.

Во-вторых, было установлено, что нахождение максимумов и минимумов требует применения достаточно сложного правила. В простейшем случае рассматривается точка на границе. Если произведение разностей между значениями частотной характеристики в данной и двух соседних точках положительно, то принимается, что в этой точке расположен локальный максимум или минимум. Для точки в средней части частотной плоскости факт размещения в ней локального максимума или минимума устанавливается в том случае, если все разности между значениями частотной характеристики в данной и восьми соседних точках имеют одинаковый знак. К сожалению, оказалось, что это простое правило не дает хороших результатов, поскольку оно очень чувствительно к погрешностям вычисления частотной характеристики. Были испробованы многие модификации этого правила. Ниже описывается модификация, дающая наилучшие результаты.

В случае когда частотная и импульсная характеристики являются действительными, поиск локальных максимумов и минимумов производится следующим образом. Прежде всего проверяются угловые точки частотной области $(\mu, \nu) = (-\pi, 0)$ и $(\mu, \nu) = (-\pi, (255/256)\pi)$. Если в любой из этих точек абсолютное значение ошибки превышает или совпадает с δ_T, то считается, что в этой точке расположен локальный максимум или минимум. Затем проверяется граница $\mu = -\pi$, $0 < \nu < (255/256)\pi$. При этом вычисляются разности между значениями частотной характеристики в данной точке границы и в двух соседних точках. Разность принимается равной нулю, если ее абсолютное зна-

чение меньше $\delta_k \cdot 10^{-5}$. Если произведение полученных разностей не является положительным, в соответствующей точке нет локального максимума или минимума. В том случае, когда произведение положительно, существуют две возможности: если в какой-то точке ошибка положительна, а обе разности отрицательны, то в этой точке размещен локальный максимум; если в какой-то точке ошибка отрицательна, а обе разности положительны, то в этой точке размещен локальный минимум. Если ошибка и обе разности имеют один и тот же знак, данная точка не учитывается. Если последовательность точек такова, что во всех промежуточных точках разности равны нулю, то рассмотренное правило применяется к ближайшим ненулевым разностям. Однако в этом случае местом размещения максимума или минимума считается точка, находящаяся в середине точечной последовательности (если требуется отыскать такую точку).

После проверки указанной границы переходят к проверке границы $-\pi \leqslant \mu \leqslant (127/128)\pi$, $\nu = 0$ и границы $-\pi \leqslant \mu \leqslant (127/128)\pi$, $\nu = (255/256)\pi$, пользуясь тем же правилом. Единственное различие состоит в том, что точки $(\mu, \nu) = ((127/128)\pi, 0)$ и $(\mu, \nu) = ((127/128)\pi, (255/256)\pi)$ не считают угловыми, поскольку действительно угловые точки принадлежат границе $\mu = \pi$, $0 \leqslant \nu \leqslant (255/256)\pi$, которая тождественна границе $\mu = -\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. Однако точки границы $\mu = \pi$ не следует включать в новое множество точек с ограничениями, поскольку ограничения для этих точек были бы тождественны ограничениям для точек границы $\mu = -\pi$.

После проверки всех границ и угловых точек следует проверить внутреннюю область $-\pi < \mu \leqslant (127/128)\pi$, $0 < \nu < (255/256)\pi$. При этом применяется следующее правило. Если все разности между значениями частотной характеристики в данной точке и в восьми соседних точках не имеют одинакового знака, эта точка не учитывается. Если в какой-то точке разности положительны, а ошибка отрицательна, то в этой точке размещен локальный минимум. Если в какой-то точке все разности отрицательны, а ошибка положительна, то в этой точке размещен локальный максимум. Точки на линии $\mu = (127/128)\pi$, $0 < \nu < (255/256)\pi$ используются в вычислениях и не считаются граничными, поскольку имеется истинная граница $\mu = \pi$. Правило для нахождения локальных максимумов и минимумов на границах полос не отличается от соответствующего правила для границы $\mu = -\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. Однако в данном случае ни одна из точек не используется как угловая, ибо для каждой такой точки имеется пара явных или неявных соседних точек.

В случае когда частотная и импульсная характеристики являются действительными и кроме того удовлетворяют условиям $h(n, m) = h(|n|, |m|)$ и $h(n, m) = h(m, n)$, правила для

нахождения локальных максимумов и минимумов в угловых точках, на границах и в пределах внутренней области частотной характеристики ничем не отличаются от только что описанных правил. Массив 256×256 отсчетов частотной характеристики заполняет область $0 \leqslant \mu \leqslant (255/256)\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$. Три точки $(\mu, \nu) = (0, 0)$, $(\mu, \nu) = ((255/256)\pi, 0)$ и $(\mu, \nu) = ((255/256)\pi, (255/256)\pi)$ считаются угловыми. Границами являются линии $0 < \mu < (255/256)\pi$, $\nu = 0$ и $\mu = (255/256)\pi$, $0 \leqslant \nu \leqslant (255/256)\pi$; кроме того, могут существовать границы полос пропускания и задерживания. Внутренние точки занимают область треугольной формы $0 < \mu < (255/256)\pi$, $0 < \nu \leqslant \mu$.

Если бы локальные минимумы и максимумы размещались только в изолированных точках частотной характеристики и если можно было бы гарантировать, что уклонения функции ошибки вдоль какой-то границы достигают максимального абсолютного значения ошибки на текущем множестве точек с ограничениями, то задача нахождения локальных максимумов и минимумов была бы намного проще. В одномерном случае эти два условия всегда удовлетворяются. В этом состоит еще одна причина того, что задача проектирования двумерных фильтров принципиально сложнее задачи проектирования одномерных фильтров. В двумерном случае локальные минимумы и максимумы могут размещаться не только в изолированных точках, но и на гребнях частотной характеристики. Функция обладает гребнем, если она имеет постоянное значение вдоль какой-то линии. Если функция ошибки данной итерации имеет гребень, то слишком много точек этого гребня выбираются как точки с локальным максимумом (или минимумом), а следовательно, как точки с ограничениями. Такая же ситуация возникает, когда имеется почти горизонтальная граница. При почти горизонтальной границе отношение сигнал/шум слишком мало, чтобы обеспечить нахождение с достаточной точностью минимумов и максимумов. Обе указанные ситуации часто возникают при проектировании фильтров нижних частот, однако ни в одном из спроектированных дифференциаторов они не отмечались. С теоретической точки зрения обсуждаемые дополнительные точки с ограничениями не создают никаких проблем. Однако с практической точки зрения эти дополнительные точки приводят к трудностям, поскольку время, которое требуется для постановки и решения задачи линейного программирования, пропорционально числу ограничений.

Пятый этап алгоритма заключается в построении графиков с выходными данными. Все графические материалы, нашедшие практическое применение, можно разделить на четыре категории:

1. Диаграммы, на которых показано местоположение точек с ограничениями, соответствующих базисным переменным для

двойственной задачи, и отдельно местоположение точек с ограничениями, соответствующих локальным максимумам и минимумам.

2. Контурные диаграммы частотной характеристики фильтра или функции ошибки.

3. Графики частотной характеристики вдоль границ полос.

4. Перспективные виды частотной характеристики или функции ошибки (или обеих функций).

Некоторые графики часто строятся при выполнении других этапов алгоритма, однако для удобства все графические построения сведены в один этап. Все они сыграли существенную роль в успешной разработке алгоритма. Они продолжают оставаться полезным выходным продуктом заключительной итерации процесса проектирования.

Шестой этап алгоритма заключается в получении данных для задачи линейного программирования и решении этой задачи. Пример формулирования задачи проектирования фильтра в виде задачи линейного программирования приведен в п. 3.1.3. Чтобы задачу в такой общей формулировке воплотить в конкретной машинной программе, требуется располагать детальными сведениями, которые существенно зависят от особенностей используемого комплекса линейного программирования. Поэтому данная операция здесь подробно не рассматривается. Путь отыскания решения задачи линейного программирования в симплексном алгоритме описан в литературе [19, 20] и не имеет значения в контексте обсуждаемой темы.

Существенной частью всякого алгоритма является правило прекращения его выполнения. Выше уже рассмотрены два условия, при которых прекращается выполнение рассматриваемого алгоритма. Вычисления прекращаются, во-первых, если δ_k (максимальное абсолютное значение ошибки на множестве точек с ограничениями для k-й итерации) не превышает δ_{k-1} или, во-вторых, если множество точек с ограничениями, соответствующих базисным переменным для k-й итерации, не отличается от множества точек для $k-1$-й итерации. Эти два условия являются достаточными, однако существует еще одно исключительно полезное условие для прекращения вычислений. Пусть N_k есть максимальное абсолютное значение ошибки по всей частотной плоскости для k-й итерации. Тогда относительная ошибка

$$E_R = \frac{N_k}{\delta_k} - 1$$

является мерой того, насколько текущая аппроксимация приблизилась к наилучшей. Малые значения E_R свидетельствуют о том, что текущая аппроксимация близка к наилучшей. Как отмечено

в п. 3.1.2, обсуждаемый алгоритм является максимизирующим методом, поскольку гарантируется увеличение от итерации к итерации δ_k (нижняя граница максимального абсолютного значения ошибки при наилучшей аппроксимации δ^*), но не гарантируется уменьшение N_k (верхняя граница δ^*). Более того, N_k обычно претерпевает заметные колебания. Однако, поскольку

$$\delta_k \leqslant \delta^* \leqslant N_k,$$

продолжение вычислений дает малый эффект, когда N_k очень близко к δ_k. Когда выполнение алгоритма не прекращалось до достижения его сходимости, были получены малые значения E_R — вплоть до 10^{-6}. Но можно получить вполне хорошие результаты, если прекратить выполнение алгоритма, когда E_R достигнет значения порядка 10^{-3}. Выбор значения E_R, при котором прекращаются вычисления, целиком определяется тем, требуется ли наилучшая аппроксимация или достаточно иметь «хорошую» аппроксимацию. Всегда имеется возможность улучшить «хорошую» аппроксимацию, определенную указанным способом, если выполнить дополнительные итерации. Кроме рассмотренных трех условий прекращения выполнения алгоритма можно также прекращать выполнение вычислений после завершения установленного числа итераций.

3.2.5. Примеры и сравнение методов

Здесь приведено несколько примеров фильтров, спроектированных с помощью алгоритмов, описанных в пп. 3.2.3 и 3.2.4. Речь идет о прямом применении линейного программирования и новом алгоритме. Рассматриваются факторы, влияющие на эффективность обоих методов. В их числе: значение того факта, что решается не двойственная, а основная задача линейного программирования; влияние числа точек с ограничениями; влияние числа независимых переменных. Кроме того, оцениваются вычислительная сложность и возможости каждого алгоритма.

Как отмечается в п. 3.1.2, если задача проектирования фильтра сформулирована как основная задача линейного программирования, то число строк в матрице ограничений намного превышает число столбцов. Если же эту задачу сформулировать в виде двойственной задачи линейного программирования, то число столбцов будет намного превышать число строк, что позволяет отыскать решение более эффективным способом. В случае прямого применения линейного программирования было установлено, что при решении основной задачи наиболее сложные фильтры, которые могли быть надежно спроектированы с помощью описанного комплекса линейного программирования,

имели импульсную характеристику размерности 7×7 (25 независимых коэффициентов импульсной характеристики). Был даже успешно спроектирован один фильтр с импульсной характеристикой размерности 9×9. Число точек с ограничениями, обработка которых не вызывает особых затруднений, составляет приблизительно 600 при полном числе ограничивающих равенств около 1200. Когда задача была переформулирована как двойственная задача линейного программирования, оказалось, что можно проектировать фильтры с импульсной характеристикой размерности 9×9 (41 независимый коэффициент) при использовании нескольких тысяч точек с ограничениями.

В табл. 3.1 представлены конкретные примеры, иллюстрирующие эту возможность. В таблице указано время, затрачиваемое в случае использования симплексного алгоритма для отыскания решения задач различной сложности, сформулированных в основной и двойственной формах. Полное время проектирования фильтра складывается из указанного времени и времени,

Таблица 3.1

Сравнение основной и двойственной форм задачи линейного программирования

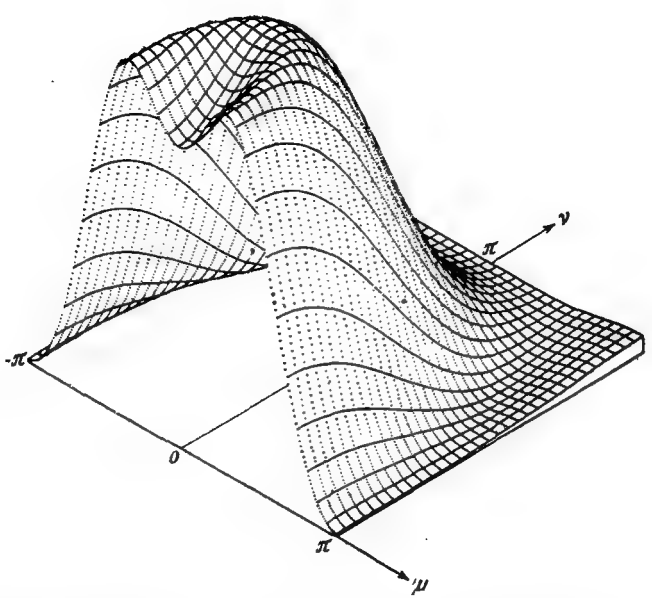
Размерность	Основная форма			Двойственная форма		
	строки	столбцы	время, мин	строки	столбцы	время, мин
5×5	378	14	1,2	14	392	0,13
7×7	918	26	16,2	26	828	1,7
9×9	1274	42	50,8	42	1316	19,5

которое требуется для получения входных данных, подготовки задачи и построения графиков с выходными данными. Представленные примеры выбраны потому, что матрицы ограничений для основной и двойственной задач имеют примерно одинаковый размер. В указанных случаях решались различные физические задачи. Очевидно, что при использовании данного комплекса линейного программирования задачу проектирования фильтра намного выгоднее формулировать в виде двойственной задачи, чем в виде основной задачи.

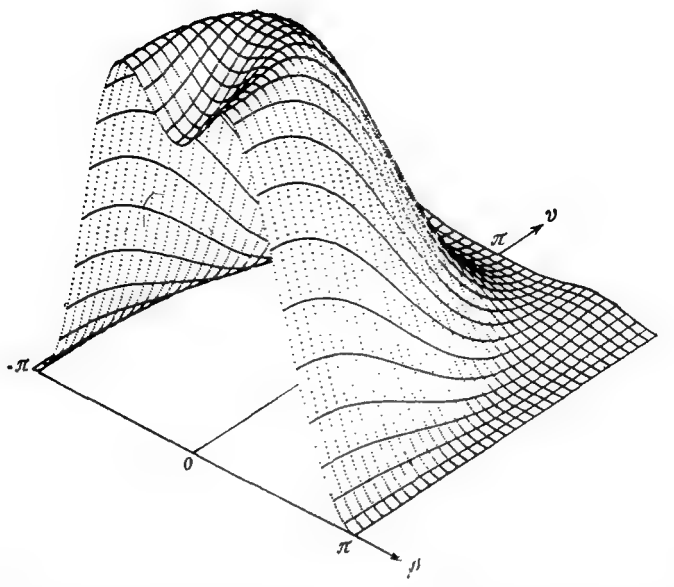
Пусть задача проектирования фильтра сформулирована как двойственная задача линейного программирования. В этом случае число столбцов в матрице ограничений равно удвоенному числу точек с ограничениями. Число строк равно числу независимых коэффициентов импульсной характеристики, увеличенному на единицу. Получаемое решение является наилучшей

аппроксимацией на данном множестве точек с ограничениями. С увеличением числа точек с ограничениями (а следовательно, числа столбцов в матрице ограничений) решение приближается к искомому решению, которое является наилучшей аппроксимацией на соответствующем подмножестве области $-\pi \leqslant \mu \leqslant \pi$, $-\pi \leqslant \nu \leqslant \pi$. Использование конечномерного множества точек с ограничениями для получения аппроксимации функции в пределах указанной области называется отбором, а ошибки, вносимые при этом, называются ошибками отбора [6].

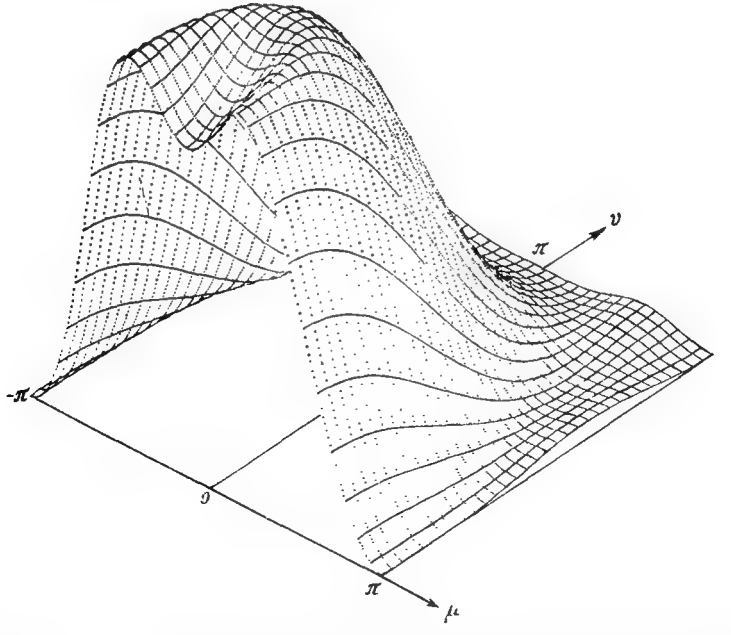
Чтобы продемонстрировать влияние отбора на результирующее решение, один и тот же фильтр нижних частот был спроектирован четырьмя различными способами. Импульсная характеристика этого фильтра имеет размерность 7×7, причем единственным принятым допущением является допущение о действительности импульсной и частотной характеристик; таким образом, всего имеется 25 независимых коэффициентов импульсной характеристики. Радиус границы полосы пропускания $R_1 = \sqrt{\mu^2 + \nu^2} = 1{,}5$; радиус границы полосы задерживания $R_2 = 2{,}5$. Местоположение точек с ограничениями было определено с помощью метода, описанного в п. 3.2.3. Первые три решения были получены путем прямого применения линейного программирова-



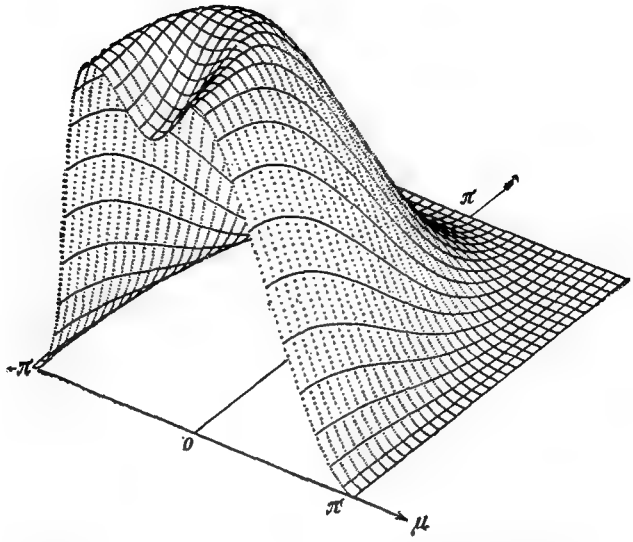
Фиг. 3.7. Влияние отбора точек с ограничениями. (Размерность решетки точек с ограничениями 7×13.)



Фиг. 3.8. Влияние отбора точек с ограничениями. (Размерность решетки точек с ограничениями 16×31.)



Фиг. 3.9. Влияние отбора точек с ограничениями. (Размерность решетки точек с ограничениями 31×61.)



Фиг. 3.10. Влияние отбора точек с ограничениями. (Новый алгоритм.)

ния. Соответствующие перспективные виды [1]) представлены на фиг. 3.7—3.9. Четвертое решение (фиг. 3.10) было получено с помощью нового алгоритма. В качестве исходной аппроксимации для этого четвертого решения было использовано первое из упомянутых трех решений.

На фиг. 3.7—3.9 видно, что выбор точек с ограничениями мало влияет на форму частотной характеристики. В других примерах были отмечены более заметные изменения формы частотной характеристики. Фиг. 3.10 ясно показывает, что функция ошибки представленной частотной характеристики остается почти неизменной на границах $-\pi \leqslant \mu \leqslant \pi$, $\nu = \pi$ и $\mu = \pm\pi$, $0 \leqslant \nu \leqslant \pi$. Именно такие ситуации, как эта, привели к разработке метода нахождения локальных максимумов и минимумов функции ошибки, описанного в п. 3.2.4. Интересно отметить, что решение на фиг. 3.10 почти удовлетворяет условиям ограничений по симметрии $h(n, m) = h(|n|, |m|)$ и $h(n, m) = h(m, n)$, хотя в этом направлении не было предпринято никаких попыток.

В табл. 3.2 приведены сравнительные данные по всем четырем решениям. Здесь дельта обозначает максимальное абсолют-

[1]) На всех фигурах, подобных фиг. 3.7—3.9, используется линейный вертикальный масштаб; данные смещены и масштаб их изменен (с учетом максимальных и минимальных значений) для заполнения всей шкалы от 0 до 1.

ное значение ошибки на множестве точек с ограничениями. Норма ошибки есть максимальное абсолютное значение ошибки на массиве 256 × 256 отсчетов частотной характеристики, вычисленных с помощью БПФ. Таблица с очевидностью показывает, что время вычислений увеличивается с увеличением числа столбцов в матрице ограничений, а дельта увеличивается с повышением плотности множества точек с ограничениями. Чем плотнее расположены точки с ограничениями, тем лучше аппроксимация (т. е. меньше норма ошибки) и тем меньше относительная

Таблица 3.2

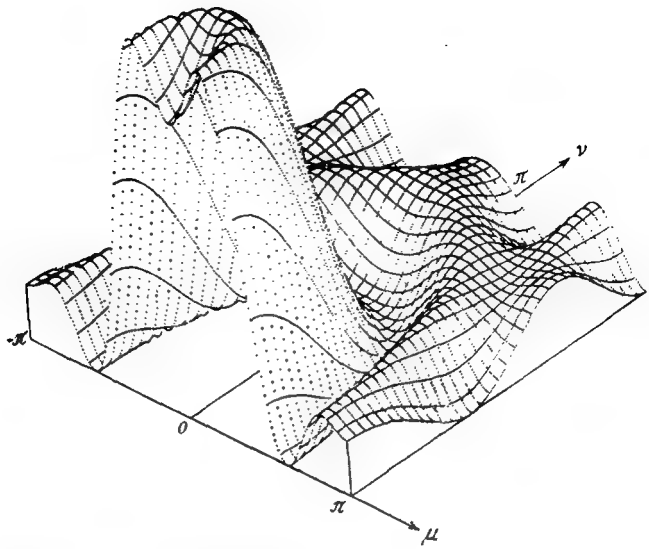
Влияние отбора точек с ограничениями

	Прямое применение линейного программирования			Новый алгоритм
Размерность решетки точек с ограничениями	7×13	16×31	31×61	
Число точек с ограничениями на границах полос	11+18	25+41	47+78	
Полное число точек с ограничениями	70	358	1311	
Дельта	0,086920	0,093992	0,094595	0,094641
Норма ошибки	0,11700	0,096651	0,094750	0,094642
Относительная ошибка	$3{,}46 \cdot 10^{-1}$	$2{,}83 \cdot 10^{-2}$	$1{,}64 \cdot 10^{-3}$	$2{,}86 \cdot 10^{-6}$
Время вычислений, мин	1,34	3,62	10,95	13,32
Число итераций				13

Примечание: относительная ошибка = норма/дельта — 1.

ошибка. Время вычислений, которое требуется для обеспечения сходимости нового алгоритма, лишь на 22% превышает время вычислений для нахождения решения с 1311 точками с ограничениями. Из 13 итераций нового алгоритма наиболее короткая имела всего 47 точек с ограничениями, а наиболее длинная — 79 точек с ограничениями. Интересно отметить, что лишь три итерации (занявшие 3,61 мин) нового алгоритма дали относительную ошибку $1{,}28269 \cdot 10^{-3}$, которая несколько меньше ошибки, полученной в наиболее объемном решении в случае прямого применения линейного программирования.

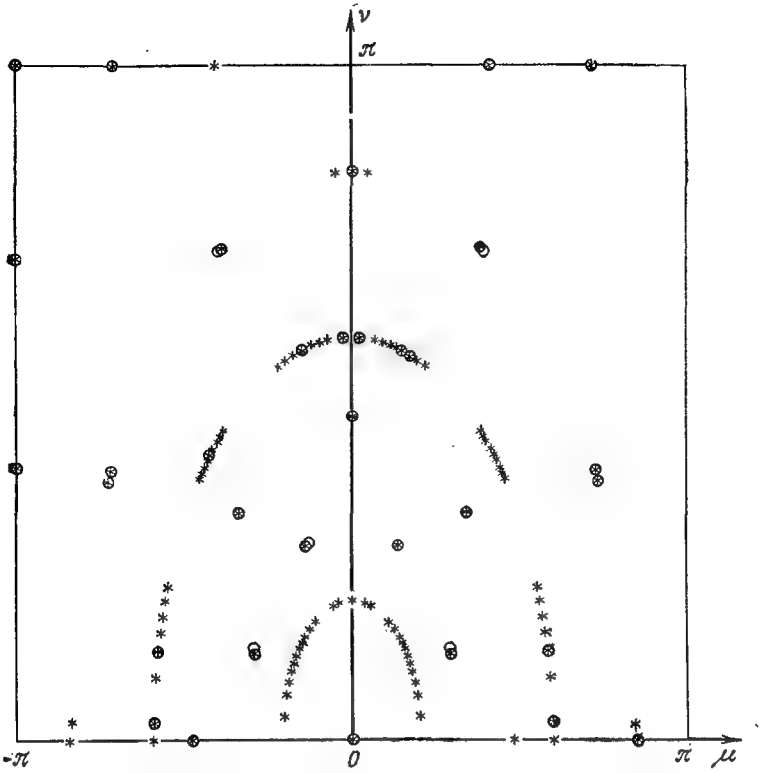
Другой фильтр, спроектированный с помощью нового алгоритма, представлен на фиг. 3.11. Его импульсная характеристика имеет размерность 9 × 9 (41 независимый коэффициент), а радиусы границ полос пропускания и задерживания равны $R_1 = 1{,}0$ и $R_2 = 1{,}5$ соответственно. Сходимость алгоритма наступала после 12 итераций с дельтой 0,115725 и нормой ошибки 0,115726. Полное время вычислений составило 31,75 мин.



Фиг. 3.11. Частотная характеристика фильтра нижних частот без ограничений по симметрии.

На фиг. 3.12 показано расположение точек, которые были бы использованы в качестве точек с ограничениями при вычислении следующей итерации. Фиг. 3.12 иллюстрирует два важных обстоятельства. Во-первых, наличие гребней частотной характеристики вынуждает находить слишком много локальных максимумов и минимумов. Поскольку имеются 42 базисные переменные для данной задачи, следует ожидать, что будет найдено приблизительно столько же локальных максимумов и минимумов. На самом же деле в рассматриваемом случае было найдено 109 локальных максимумов и минимумов. Выше было отмечено, что это обстоятельство не создает трудностей теоретического плана, хотя и приводит к увеличению полного времени вычислений. Во-вторых, важно отметить, что точки с ограничениями, соответствующие базисным переменным, в результирующем решении, а следовательно, и в предшествующих итерациях часто появляются парами (иногда тройками). Это явление наблюдалось почти во всех примерах проектирования фильтров. Объяснить это явление не удалось.

Как и в последнем примере, в данном случае не предпринималось никаких попыток к тому, чтобы решение удовлетворяло условиям ограничений по симметрии $h(n, m) = h(|n|, |m|)$ и

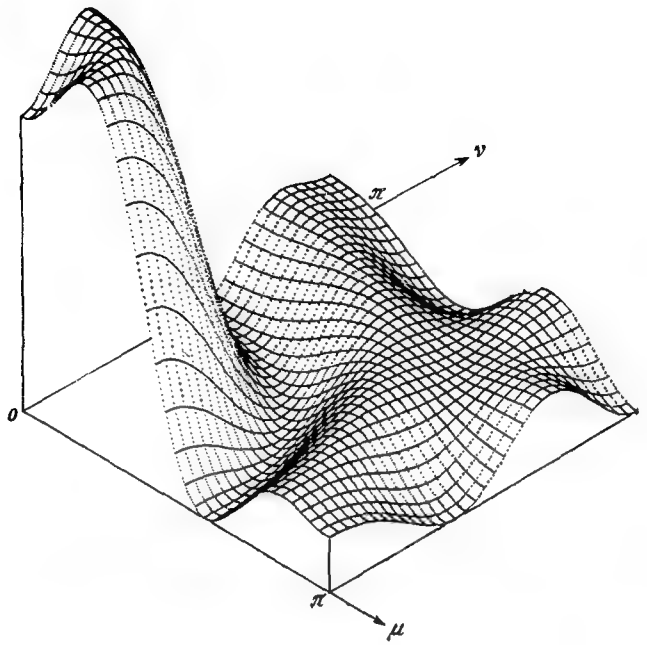


Фиг. 3.12. Размещение точек с ограничениями для частотной характеристики на фиг. 3.11; ⊛ — локальные максимумы и минимумы, ⊙ — базисные переменные.

$h(n, m) = h(m, n)$. Была поставлена цель просто определить наилучшую аппроксимацию функции

$$F(\mu, \nu) = \begin{cases} 1 \text{ при } \sqrt{\mu^2 + \nu^2} \leqslant 1{,}0, \\ 0 \text{ при } \sqrt{\mu^2 + \nu^2} \geqslant 1{,}5. \end{cases}$$

Тем не менее оказалось, что результирующее решение почти полностью удовлетворяет условиям ограничений по симметрии. Не имеется доказательства того, что это должно всегда происходить в случае фильтров (нижних частот), подобных рассматриваемым. Однако этот факт отмечался и в ряде других примеров. Если ограничения по симметрии установить заранее, то будет найдено решение, которое иллюстрирует фиг. 3.13. При любом практическом подходе можно считать, что это решение тождественно решению, показанному на фиг. 3.11. В данном



Фиг. 3.13. Частотная характеристика фильтра нижних частот с ограничениями по симметрии.

случае имеется всего 15 независимых коэффициентов импульсной характеристики. Сходимость алгоритма наступила после 18 итераций с дельтой 0,115726 и нормой ошибки 0,115727. Полное время вычислений составило 17,45 мин. Интересно отметить, что при наложении ограничений по симметрии нахождение оптимального решения потребовало шести дополнительных итераций алгоритма, однако время вычислений уменьшилось на 14,3 мин.

Следующие четыре примера фильтров, спроектированных с помощью нового алгоритма, иллюстрируют влияние размерности импульсной характеристики на результирующую частотную характеристику. В качестве идеального фильтра был выбран двумерный дифференциатор с частотной характеристикой вида

$$F(\mu, \nu) = \frac{1}{2\pi^2}(\mu^2 + \nu^2).$$

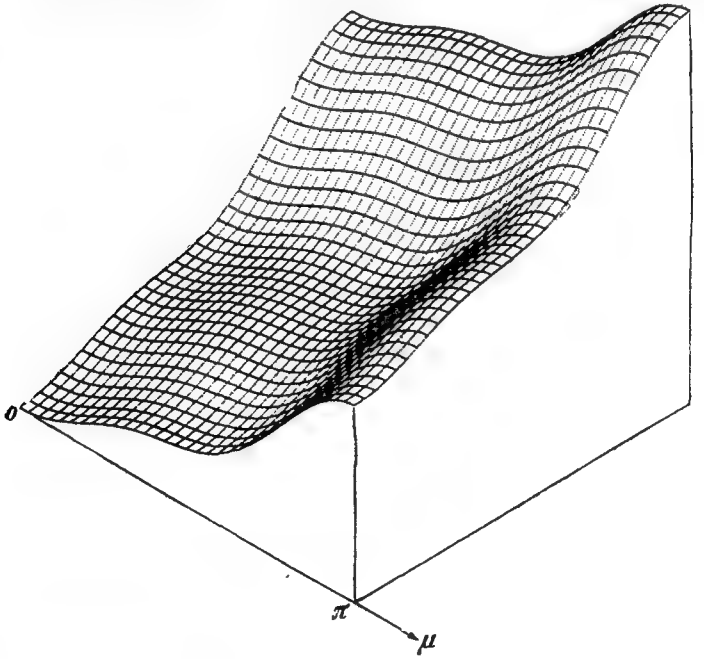
Было наложено ограничение по симметрии $h(n, m) = h(|n|, |m|)$ и $h(n, m) = h(m, n)$. На фиг. 3.14—3.17 показаны частотные характеристики каждого фильтра в одном квадранте. Табл. 3.3 содержит сравнительные данные по этим фильтрам. В каждом

Таблица 3.3

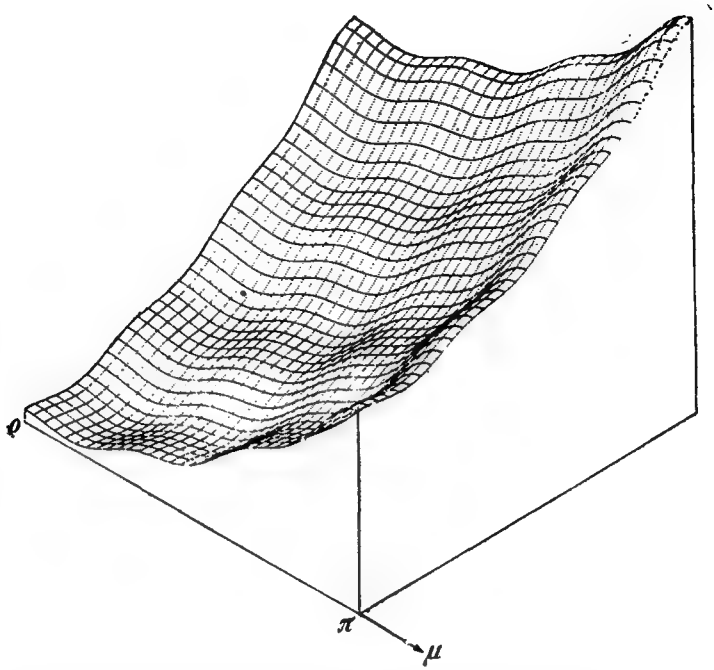
Примеры дифференциаторов

	Фиг. 3.14	Фиг. 3.15	Фиг. 3.16	Фиг. 3.17
Размерность импульсной характеристики	7×7	13×13	15×15	17×17
Число независимых коэффициентов	10	28	36	45
Дельта	0,055345	0,026919	0,022752	0,019636
Норма ошибки	0,055359	0,026933	0,022767	0,019649
Относительная ошибка	$2{,}55 \cdot 10^{-4}$	$5{,}09 \cdot 10^{-4}$	$6{,}28 \cdot 10^{-4}$	$7{,}05 \cdot 10^{-4}$
Время вычислений, мин	8,68	33,55	38,55	71,27
Число итераций	9	24	23	32
Среднее время вычисления одной итерации, мин	0,964	1,398	1,675	2,227

случае алгоритм вычислялся до полного схождения. Полученные значения относительных ошибок (порядка 10^{-4}) на один или два порядка превышают значения относительных ошибок, обычно



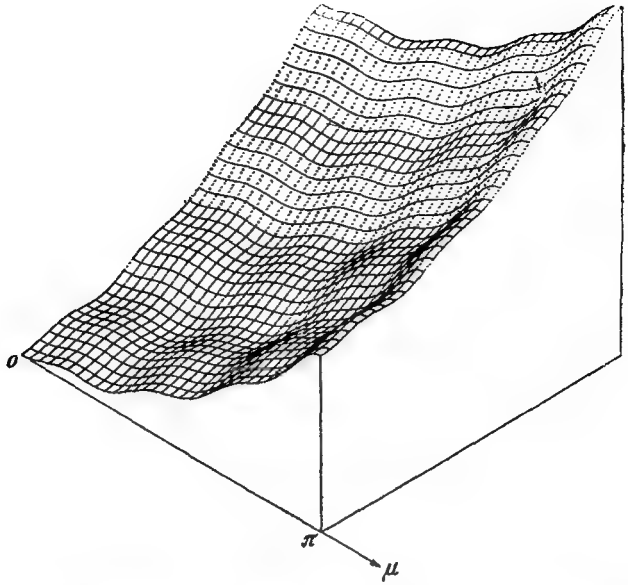
Фиг. 3.14. Аппроксимация частотной характеристики идеального дифференциатора. (Размерность импульсной характеристики 7 × 7.)



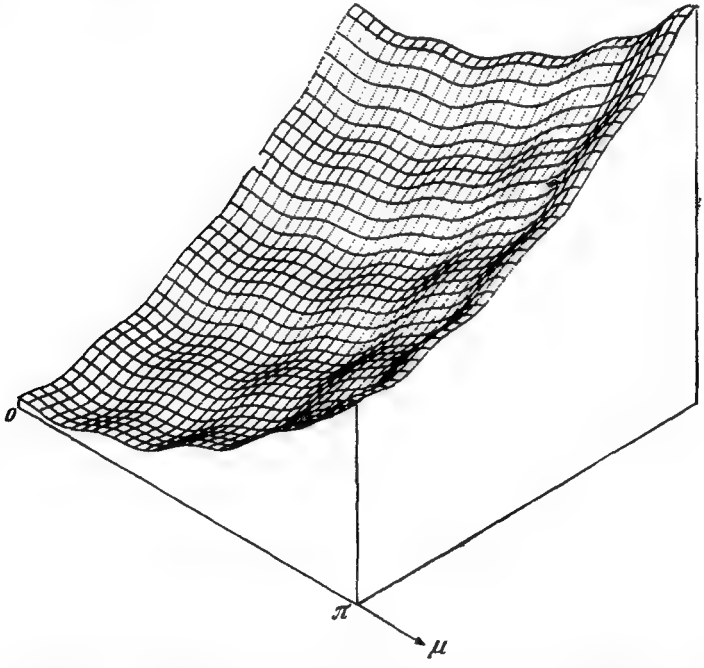
Фиг. 3.15. Аппроксимация частотной характеристики идеального дифференциатора. (Размерность импульсной характеристики 13 × 13.)

получаемых при проектировании фильтров нижних частот. Объяснение этому не найдено. Также не имеет объяснения тот факт, что функция ошибки никогда не имела гребней. В каждой из итераций, выполненных во всех примерах, все локальные максимумы и минимумы располагались в изолированных точках функции ошибки. Интересно отметить, что в примере с импульсной характеристикой размерности 13 × 13 потребовалось иметь на одну итерацию больше, чем в примере с импульсной характеристикой размерности 15 × 15. Однако в целом полное время вычислений и время вычисления одной итерации увеличиваются с увеличением размерности импульсной характеристики. Переход от фильтра с 10 независимыми коэффициентами к фильтру с 45 независимыми коэффициентами обеспечил уменьшение дельты в 2,82 раза. При этом полное время вычислений увеличилось в 8,22 раза, а время вычисления одной итерации — в 2,32 раза.

Приведенные примеры свидетельствуют о том, что линейное программирование можно с успехом использовать для проектирования двумерных нерекурсивных цифровых фильтров. Теоретически симплексный алгоритм может дать решение для любой



Фиг. 3.16. Аппроксимация частотной характеристики идеального дифференциатора. (Размерность импульсной характеристики 15 × 15.)



Фиг. 3.17. Аппроксимация частотной характеристики идеального дифференциатора. (Размерность импульсной характеристики 17 × 17.)

задачи линейного программирования без каких-либо исключений. На практике, однако, приходится беспокоиться о вычислительной сложности алгоритма. На сложность алгоритма указывает схема допусков, которой неизбежно приходится пользоваться при реализации алгоритма вследствие отсутствия арифметических устройств с бесконечной точностью. В основном требуется использовать допуски двух видов: по точности и по значимости. Допуски по точности позволяют проводить определенную самопроверку, что обеспечивает возможность приостановки вычислений на ранних стадиях в случае обнаружения существенной потери точности. Допуски по значимости налагаются таким образом, что значения, очень близкие к нулю, при выполнении вычислений заменяются нулями. Эти допуски также могут стать причиной досрочной приостановки вычислений, если будут отсутствовать «ненулевые» элементы матрицы ограничений, требуемые для продолжения вычислений.

Симплексный алгоритм по своей сути является итеративным, и это гарантирует уменьшение (или увеличение в случае задачи на максимизацию) целевой функции с каждой следующей итерацией. Более того, имеется гарантия того, что алгоритм обязательно сойдется и будет найдено решение при выполнении конечного числа итераций. Однако природа алгоритма такова, что совершенно невозможно предсказать число итераций, которые фактически потребуются для решения конкретной задачи. Две задачи могут иметь совершенно одинаковое число строк и столбцов в матрице ограничений, однако одна из них может потребовать больше времени для решения, чем другая. В случае проектирования фильтров это выражается следующим образом. Тот факт, что фильтр данной размерности может быть спроектирован за определенное время, еще не доказывает возможности проектирования другого фильтра той же размерности и с тем же числом точек с ограничениями.

Аналогичные замечания, очевидно, можно сделать в отношении вычислительной сложности нового алгоритма, поскольку при выполнении каждой итерации этого алгоритма приходится решать задачу линейного программирования. Сложность усугубляется тем обстоятельством, что в каждой итерации требуется находить локальные максимумы и минимумы функции ошибки. Выше уже обсуждались проблемы, возникающие при нахождении этих точек. Очевидно, что фактическое решение, получаемое в результате каждой итерации, зависит от выбора точек с ограничениями, однако очень трудно проследить такую зависимость. Как и прежде, в данном случае невозможно предсказать число итераций, которые потребуется вычислить. Максимальное абсолютное значение ошибки на множестве точек с ограничениями при переходе от одной итерации к следующей

возрастает, однако такое приращение может оказаться очень малым. В такой ситуации нельзя прекращать вычисление алгоритма, поскольку через несколько итераций величина приращения может вновь оказаться значительной. При относительно плотном множестве точек с ограничениями время для схождения нового алгоритма может оказаться сравнимым или меньшим времени, которое потребовалось бы для проектирования того же фильтра методом прямого применения линейного программирования. Однако иногда случается так, что новый алгоритм требует большего времени вычислений, давая лишь незначительное улучшение аппроксимации. Создается впечатление, что лучше всего пользоваться новым алгоритмом следующим образом: надо принять решение о допустимом значении относительной ошибки и прекратить вычисление после достижения этого значения. Если требуются более качественные аппроксимации, то их можно получить путем вычисления дополнительных итераций. В этом состоит важное преимущество нового алгоритма перед методом прямого применения линейного программирования, при котором в случае получения недостаточно хорошей аппроксимации требуется повторить все вычисления.

Новый алгоритм имеет и другие преимущества перед прямым применением линейного программирования. Во-первых, новый алгоритм предполагает вычисление наилучшей аппроксимации требуемой частотной характеристики на множестве 256×256 точек частотной области; напротив, задача линейного программирования решается только на множестве точек с ограничениями. В этой связи следует сделать два замечания. Первое заключается в том, что размерность 256×256 есть практическое, а не теоретическое ограничение. Увеличив объем БПФ, можно получить более качественные аппроксимации на более плотных множествах. Однако концепция алгоритма остается неизменной. Второе замечание заключается в том, что решить задачу линейного программирования с $2 \times 256 \times 256 = 131\,072$ ограничивающими уравнениями просто невозможно. Новый алгоритм имеет еще одно преимущество: он позволяет проектировать фильтры, размерность которых несколько превышает размерность фильтров, проектируемых другими методами. Это объясняется тем, что алгоритм фактически позволяет расчленить объемную задачу линейного программирования на ряд небольших задач. Поскольку время вычисления одной итерации и число итераций увеличиваются с возрастанием размерности фильтра, указанное преимущество, к сожалению, нельзя считать существенным, за исключением того случая, когда время вычислений не имеет никакого значения. Например, прямое применение линейного программирования ни разу не позволило успешно спроектировать фильтр с импульсной характеристикой

размерности 11×11 (61 независимый коэффициент). Была предпринята попытка спроектировать этот фильтр с помощью нового алгоритма. Однако после нескольких итераций вычисления пришлось прекратить, поскольку выяснилось, что вычисление одной итерации занимало приблизительно шесть минут.

3.3. Краткое содержание и выводы

Рассмотрены четыре метода проектирования двумерных нерекурсивных цифровых фильтров. Первые три — применение функций окна, частотная дискретизация и прямое применение линейного программирования — являются прямыми обобщениями соответствующих методов проектирования одномерных фильтров. Только третий из указанных методов потенциально позволяет проектировать оптимальные фильтры. Четвертый метод — новый алгоритм, который был специально разработан для решения задач проектирования двумерных фильтров. Этот метод превосходит метод прямого применения линейного программирования, поскольку: а) «хорошие» аппроксимации обычно вычисляются быстро, причем эти аппроксимации всегда можно улучшить просто путем вычисления дополнительных итераций; б) наилучшие аппроксимации удается получить даже при чрезвычайно высокой плотности множества точек; в) обеспечивается возможность расчленить объемную задачу линейного программирования на ряд небольших задач. К сожалению, новый алгоритм оказался не таким эффективным с вычислительной точки зрения, как ожидалось.

Один из самых важных и глубоких выводов, сделанных в результате проведенного исследования, состоит в том, что задача проектирования двумерных фильтров в ее общем виде существенно сложнее задачи проектирования одномерных фильтров. Причин для этого много. Наиболее важной, по-видимому, является та, что множество функций, используемых для вычисления аппроксимации, не удовлетворяет условию Хаара (п. 3.1.2). Вследствие этого невозможно написать общую интерполяционную формулу, аналогичную одномерной интерполяционной формуле Лагранжа [27—29, 32]. Более того, поскольку множество функций не удовлетворяет условию Хаара, вопрос об единственности наилучшей аппроксимации становится исключительно сложным [33]; в одномерном случае наилучшая аппроксимация является единственной. Наконец, теорема чередования, которая является одной из важнейших теорем теории одномерной аппроксимации, неприменима к многомерному случаю. Это объясняется тем фактом, что функции двух и более переменных обычно не удовлетворяют условию Хаара.

Двумерная задача труднее одномерной еще и потому, что в двумерном случае обычно приходится иметь дело с большим массивом данных. Поэтому в двумерном случае значительно возрастают требования к объему памяти и время вычислений. В качестве примера рассмотрим задачу вычисления частотной характеристики фильтра. В одномерном случае действует следующее практическое правило: частотную характеристику следует вычислять на множестве точек, число которых в 20 раз превышает число отсчетов импульсной характеристики фильтра. Таким образом, если импульсная характеристика фильтра содержит 100 отсчетов, то частотную характеристику следует вычислять на множестве 2000 точек. Во всех примерах данной работы частотная характеристика фильтров вычислялась на множестве 256×256 точек, т. е. общее число отсчетов частотной характеристики составляет 65 536.

Существует ряд алгоритмов проектирования одномерных фильтров, основанных на отыскании экстремумов, т. е. локальных максимумов и минимумов функции ошибки. Эта операция также является важнейшей составляющей нового алгоритма, который был специально разработан для проектирования двумерных фильтров. Следует указать, что некоторые явления возникают только в двумерном случае, и они значительно усложняют отыскание экстремумов в сравнении с одномерным случаем. Первое такое явление — наличие гребней в функции ошибки. Это явление, очевидно, не может существовать в одномерном случае. Как отмечается в п. 3.2.4, оно существенно усложняет задачу отыскания экстремумов по сравнению с другими возможными случаями. Второе явление, влияющее на поиск экстремумов, состоит в существовании почти горизонтальных границ (имеются в виду границы полос или границы $\mu = \pm\pi$ и $\nu = \pi$) в функции ошибки. Отыскание локальных максимумов и минимумов на таких границах представляет собой одномерную задачу. Однако в данном случае задача сложнее, чем в одномерном случае, когда уклонение функции ошибки по крайней мере равно максимальному абсолютному значению ошибки на последнем множестве точек с ограничениями. В двумерном случае ошибка вдоль таких границ может сохранять почти постоянное значение. При этом очень трудно отыскать локальные максимумы и минимумы вдоль таких границ.

Графическое представление частотной характеристики двумерного фильтра значительно сложнее графического представления частотной характеристики одномерного фильтра. В одномерном случае достаточно пользоваться простым графиком. В двумерном случае частотная характеристика имеет вид поверхности, поэтому важно уметь строить проекции такой поверхности. Если частотная характеристика фильтра не обладает

высокой степенью симметрии, может возникнуть необходимость отображения указанной поверхности с использованием нескольких различных точек проекции. Однако и в этом случае часто оказываются полезными контурные диаграммы частотной характеристики и (или) функции ошибки, а также диаграммы с указанием местоположения локальных максимумов и минимумов функции ошибки. Перечисленные формы графического представления значительно сложнее и требуют больших затрат времени, чем простые графики, позволяющие дать полную информацию о частотной характеристике одномерного фильтра.

Последняя причина, по которой задача проектирования двумерного фильтра сложнее задачи проектирования одномерного фильтра, связана исключительно с линейным программированием. При любом заданном числе независимых коэффициентов импульсной характеристики в двумерной задаче почти всегда используется больше точек с ограничениями, чем в одномерной задаче. Это утверждение справедливо в том случае, когда метод прямого применения линейного программирования выбирается только потому, что плотное множество точек с ограничениями должно покрыть не одну линию, а целую область. Если задача проектирования одномерного фильтра решается с использованием итеративных алгоритмов, число точек с ограничениями, требующихся для выполнения каждой итерации, равно числу независимых коэффициентов импульсной характеристики, увеличенному на единицу; указанное число не меняется от итерации к итерации. В новом алгоритме проектирования двумерных фильтров число точек с ограничениями для вычисления каждой итерации равно сумме увеличенного на единицу числа независимых коэффициентов импульсной характеристики и числа локальных максимумов и минимумов текущей функции ошибки. Второе из этих двух чисел в большинстве случаев меняется в значительных пределах (исключение составляют спроектированные автором дифференциаторы), что зависит от наличия или отсутствия гребней и горизонтальных границ полос в функции ошибки. Увеличение числа точек с ограничениями обычно приводит к увеличению времени нахождения решения.

Сразу же возникает вопрос о том, насколько задача проектирования трехмерного фильтра окажется сложнее задачи проектирования двумерного фильтра. Все теоретические положения в п. 3.1.2, относящиеся к проектированию фильтров, были сформулированы применительно к задаче проектирования двумерных фильтров. Это было сделано исключительно ради удобства изложения. Все полученные результаты, применимые к двумерному случаю, в равной степени применимы к трех-, четырех- и n-мерному случаям. Как и в двумерном случае, множество функций, используемое для вычисления аппроксимации, в об-

щем случае не будет удовлетворять условию Хаара. Возникнут трудности, связанные с гребнями и равноуровневыми границами полос (хотя и в более обобщенном смысле), а n-мерное z-преобразование выдвинет проблемы, неизвестные в одномерном случае. Переход от двумерного к n-мерному случаю не должен привести к усугублению трудностей в этих вопросах. Однако имеется ряд проблем, решение которых существенно затруднится. Во-первых, будет чрезвычайно трудно или даже совершенно невозможно дать графическое представление частотной характеристики спроектированного фильтра в сколько-нибудь удобной форме. Во-вторых, обращение с большими массивами данных многократно усложнит трудности, связанные с обеспечением требуемых объема памяти и времени вычислений. Наконец, в случае применения линейного программирования существенно возрастут трудности, обусловленные большим числом точек с ограничениями при заданной размерности импульсной характеристики.

От автора. Представленная работа была выполнена в Линкольновской лаборатории Массачусетского технологического института и финансировалась министерством военно-воздушных сил. Автор выражает признательность д-ру Голду, д-ру Хофстеттеру, проф. Хуангу, проф. Штёлину, проф. Оппенгейму и проф. Штейглицу за помощь, оказанную ими при выполнении этой работы. Данная глава основана на материалах диссертации «Двумерные нерекурсивные цифровые фильтры», представленной на рассмотрение кафедры электротехники Массачусетского технологического института 4 мая 1973 г. на соискание степени доктора философии.

Литература

1. *Helms H. D.*, IEEE Trans. Audio Elektroacoust., **AU-16**, 336 (1968).
2. *Rabiner L. R.*, IEEE Trans. Commun. Technol., **COM-19**, 188 (1971).
3. *Kaiser J. F.*, Digital Filters, в книге System Analysis by Digital Computer (eds. Kuo F. F., Kaiser J. F.), J. Wiley, New York, 1966. (Имеется перевод: см. [13].)
4. *Rabiner L. R., Gold B., McGonegal C. A.*, IEEE Trans. Audio Electroacoust., **AU-18**, 83 (1970).
5. *Rabiner L. R.*, IEEE Trans. Audio Electroacoust., **AU-20**, 280 (1972).
6. *Cheney E. W.*, Introduction to Approximation Theory, McGraw-Hill, New York, 1966.
7. *Herrmann O.*, Electron. Letters, **6** (1970).
8. *Hofstetter E. M., Oppenheim A. V., Siegal J.*, A New Technique for the Design of Nonrecursive Digital Filters, Proc. Fifth Annual Princeton Conference on Information Sciences and Systems, 1971.
9. *Parks T. W., McClellan J. H.*, IEEE Trans. Circuit Theory, **CT-19**, 189 (1972).
10. *Huang T. S.*, IEEE Trans. Audio Electroacoust., **AU-20**, 158 (1972).

11. *Shanks T. L., Treitel S., Justice J. H.*, IEEE Trans. Audio Electroacoust., **AU-20**, 115 (1972).
12. *Jury E. I.*, Theory and Application of the Z-Transform Method, J. Wiley, New York, 1964.
13. *Gold B., Rader C. M.*, Digital Processing of Signals, McGraw-Hill, New York, 1969. (Имеется перевод: *Голд Б., Рэйдер Ч.* Цифровая обработка сигналов. — М.: Советское радио, 1973.)
14. *Hanscomb D. C.*, Functions of Many Variables, Methods of Numerical Approximation, ed. Hanscomb D. C., Pergamon Press, New York, 1966.
15. *Rice J. R.*, The Approximation of Functions, Vol. II, Addison — Wesley, Reading, Mass., 1969.
16. *Rice J. R.*, The Approximation of Functions, Vol. I, Addison — Wesley, Reading, Mass., 1964.
17. *Ремез Е. Я.*, Общие вычислительные методы чебышевского приближения, АН УССР, Киев, 1957.
18. *Stiefel E. L.*, Numerical Methods of Tchebycheff Approximation, On Numerical Approximation (ed. Langer R. E.), The University of Wisconsin Press, 1959.
19. *Garvin W. W.*, Introduction to Linear Programming, McGraw-Hill, New York, 1960.
20. *Spivey W. A., Thrall R. M.*, Linear Optimization, Holt, Rinehart and Winston, New York, 1970.
21. *Huang T. S.*, IEEE Trans. Audio Electroacoust., **AU-20**, 88 (1972).
22. *Papoulis A.*, Systems and Trasforms with Applications in Optics, McGraw-Hill, New York, 1968. (Имеется перевод: *Папулис А.* Теория систем и преобразований в оптике. — М.: Мир, 1971.)
23. *Rabiner L. R.*, Processing of Two-Dimensional Signals, Digital Signal Processing by Rabiner L. R., Gold B. (unpublished).
24. *Hu J. V., Rabiner L. R.*, IEEE Trans. Audio Electroacoust., **AU-20**, 249 (1972).
25. *Hu J. V.*, Frequency Sampling Design of Two-Dimensional Finite Impulse Response Digital Filters, M. I. T. S. M. Thesis, E. E. Dept., 1972.
26. IBM System/360 Scientific Subroutine Package (360A-CM-03 X), Version III, Programmer's Manual.
27. *Milne W. E., Arntzen W., Reynolds N., Wheelock J.*, Mathematics for Digital Computers, I, Multivariate Interpolation WADC Technical Report 57—556 (1958) ASTIA Document № AD 131033.
28. *Thacher H. C., Jr., Milne W. E.*, J. Soc. Indust. Appl. Math., **8**, 33 (1960).
29. *Guenther R. B., Roetman E. L.*, Math. of Computation, **24**, 517 (1970).
30. *Brenner N. M.*, Three Fortran Programs that Perform the Cooley-Tukey Fourier Transform, Technical Note 1967-2, Lincoln Laboratory, M. I. T. (July 1967).
31. *Rabiner L .R., Schafer R. W., Rader C. M.*, IEEE Trans. Audio Electroacoust., **AU-17**, 86 (1969).
32. *Hamming R. W.*, Numerical Methods for Scientists and Engineers, McGraw-Hill, New York, 1962. (Имеется перевод: *Хемминг Р. В.* Численные методы для научных работников и инженеров. — М.: Наука, 1968.)
33. *Rivlin T. J., Shapiro H. S.*, Comm. Pure Appl. Math., **13**, 35 (1960).

4. Двумерная рекурсивная фильтрация

Р. Рид, Дж. Шэнкс, С. Трейтел

4.1. Введение

Определение рекурсивного фильтра через понятия z-преобразования

Двумерный цифровой рекурсивный фильтр характеризуется двумерной передаточной функцией

$$H(z_1, z_2) = \frac{A(z_1, z_2)}{B(z_1, z_2)} = \frac{\sum_{i=1}^{M_a}\sum_{j=1}^{N_a} a_{ij} z_1^{i-1} z_2^{j-1}}{\sum_{k=1}^{M_b}\sum_{l=1}^{M_b} b_{kl} z_1^{k-1} z_2^{l-1}}. \tag{4.1}$$

Здесь a_{ij} и b_{kl} — константы. Без ущерба для общности можно полагать, что $b_{11} = 1$. Пусть $D(z_1, z_2)$ есть z-преобразование входного массива; тогда z-преобразование массива на выходе фильтра с передаточной функцией $H(z_1, z_2)$ выражается в виде

$$R(z_1, z_2) = H(z_1, z_2) \cdot D(z_1, z_2). \tag{4.2}$$

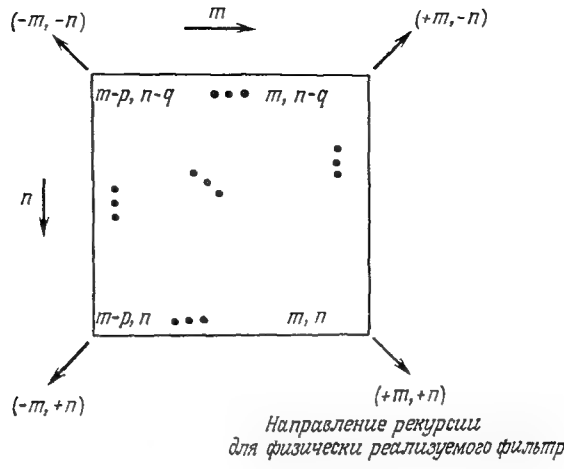
Решая уравнение относительно $R(z_1, z_2)$, получаем

$$R(z_1, z_2) = \left[\sum_{i=1}^{M_a}\sum_{j=1}^{N_a} a_{ij} z_1^{i-1} z_2^{j-1}\right] D(z_1, z_2) - \left[\sum_{\substack{k=1 \\ k \cdot l \neq 1}}^{M_b}\sum_{l=1}^{N_b} b_{kl} z_1^{k-1} z_2^{l-1}\right] R(z_1, z_2), \tag{4.3}$$

поскольку $b_{11} = 1$; здесь $R(z_1, z_2)$ — z-преобразование массива на выходе фильтра. Значение каждого выходного отсчета r_{mn} можно выразить через значения уже вычисленных отсчетов r.

Физическая реализуемость

Согласно принятой здесь системе обозначений, z_1 и z_2 обозначают единичные задержки по горизонтали и вертикали соответственно. Выражение (4.3) описывает фильтр с рекурсией в направлении положительных m и n. Можно образовать рекурсивный фильтр с рекурсией в любом из трех других диагональных направлений, если произвести нормирование относительно



Фиг. 4.1. Четыре направления рекурсии (по Хуангу).

коэффициентов соответствующего угла массива B. Такую возможность иллюстрирует фиг. 4.1 (где для упрощения принято $M_a = M_b = p - 1$ и $N_a = N_b = q - 1$). Рекурсивный фильтр с рекурсией в направлении положительных m и n называется «физически реализуемым». В случае отсутствия каких-либо оговорок будем считать, что (4.1) описывает физически реализуемый рекурсивный фильтр.

Импульсная характеристика

Импульсная характеристика рекурсивного двумерного фильтра представляет собой сигнал на его выходе в том случае, когда входной сигнал задается матрицей U, элемент u_{11} которой равен единице, а остальные элементы равны нулю. Можно также считать, что импульсная характеристика рекурсивного цифрового фильтра есть разложение в двумерный степенной ряд функции (4.1).

Применение рекурсивного цифрового фильтра для обработки двумерного массива данных возможно при соблюдении двух условий: во-первых, фильтр должен быть правильно спроектирован и, во-вторых, должен быть обеспечен ввод данных в фильтр. Проектирование рекурсивных фильтров требует решения двух основных проблем: проблемы аппроксимации и проблемы устойчивости. Первая заключается в выборе таких коэффициентов фильтра, чтобы его характеристики удовлетворяли тому или иному выбранному критерию. Обычно требуется обеспечить заданную частотную характеристику. Однако в некото-

рых случаях требуется выбрать такие коэффициенты цифрового фильтра, чтобы его импульсная характеристика аппроксимировала заданную импульсную характеристику.

Абсолютная устойчивость

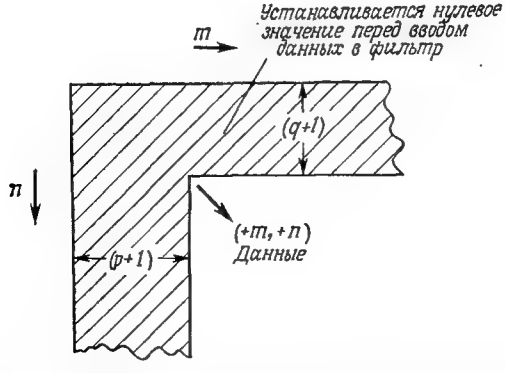
Проблема устойчивости сводится к наложению ограничений на коэффициенты фильтра таким образом, чтобы его импульсная характеристика удовлетворяла условию

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}|h_{mn}|<\infty, \tag{4.4}$$

где h_{mn} — импульсная характеристика фильтра. В неустойчивом фильтре любой шум, включая ошибки округления при выполнении вычислений, появится на его выходе и будет усилен.

Начальные условия

В случае физически реализуемого фильтра часто используют нулевые начальные условия, устанавливаемые перед вводом данных в фильтр (см. фиг. 4.2, где для упрощения принято снова $M_a = M_b = p - 1$ и $N_a = N_b = q - 1$). Результат воздействия двумерного фильтра на данные отдаленно напоминает реакцию одномерного фильтра на ступенчатую функцию. Если $|h_{mn}| \approx 0$ при $m, n > L$, причем L мало по сравнению с длиной входного и выходного массивов по обоим направлениям, то в самом начале функционирования фильтра эффект ступенчатого возмущения будет проявляться в минимальной степени. Таким образом, в некоторых случаях для обеспечения устойчивости недостаточно удовлетворить только условию (4.4), а требуется иметь быстро спадающую импульсную характеристику.



Фиг. 4.2. Начальные условия для физически реализуемого фильтра (по Хуангу).

4.2. Устойчивость

Устойчивость двумерного рекурсивного фильтра определяют коэффициенты знаменателя $B(z_1, z_2)$ в выражении (4.1). Обобщенный критерий устойчивости дает теорема Шэнкса [10].

4.2.1. Теорема 1 (теорема Шэнкса)

Пусть $B(z_1, z_2)$ есть полином относительно (z_1, z_2); для того чтобы коэффициенты разложения $1/B(z_1, z_2)$ по положительным степеням z_1 и z_2 сходились в абсолютном смысле, необходимо и достаточно, чтобы полином $B(z_1, z_2)$ не принимал нулевых значений при $|z_1|$ и $|z_2|$ одновременно меньших или равных единице.

В другой формулировке теорема утверждает, что если существуют любые значения z_1 и z_2, при которых $B(z_1, z_2)$ равен нулю, причем z_1 и z_2 одновременно меньше или равны единице по величине, то фильтр с передаточной функцией $1/B(z_1, z_2)$ является неустойчивым. Если таких значений не существует, фильтр является устойчивым.

Доказательство [1]) [1]. Пусть

$$H(z_1, z_2) = \frac{A(z_1, z_2)}{B(z_1, z_2)} = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} h_{mn} z_1^m z_2^n. \tag{4.5}$$

Требуется доказать, 1) что

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} |h_{mn}| < \infty,$$

если $H(z_1, z_2)$ является аналитической функцией в области $D = \{(z_1, z_2);\ |z_1| \leqslant 1 \cap |z_2| \leqslant 1\}$, и 2) что указанное неравенство справедливо только в том случае, если $H(z_1, z_2)$ является аналитической функцией в указанной области.

Первая часть доказательства. Если $H(z_1, z_2)$ есть аналитическая функция в области D, то можно найти такое $\varepsilon > 0$, что $H(z_1, z_2)$ будет аналитической функцией и в области

$$D_1 = \{(z_1, z_2);\ |z_1| < 1 + \varepsilon \cap |z_2| < 1 + \varepsilon\},$$

из чего следует, что

$$\sum_m \sum_n h_{mn} z_1^m z_2^n$$

[1]) Это доказательство справедливо для случая, когда $B(z_1, z_2)$ является полиномом относительно z_1 и z_2. Доказательство теоремы для случая, когда $B(z_1, z_2)$ является степенным рядом, т. е. содержит бесконечное число коэффициентов, дано Джастисом и Шэнксом [9].

сходится в области D_1 в абсолютном смысле. Поэтому

$$\sum_m \sum_n |h_{mn}| < \infty.$$

Вторая часть доказательства. Если в M проверках

$$\sum_m \sum_n |h_{mn}| < \infty,$$

то

$$\sum_m \sum_n h_{mn} z_1^m z_2^n$$

сходится в абсолютном смысле в области D; последнее означает, что $H(z_1, z_2)$ является аналитической функцией в области D (что и требовалось доказать).

Корневые диаграммы

Чтобы воспользоваться данной теоремой для проверки устойчивости, вообще говоря, требуется определить весь континуум значений (z_1, z_2), при которых $B(z_1, z_2)$ принимает нулевое значение. Область этих нулевых значений можно оценить следующим образом. Можно придавать переменной z_1 различные значения и находить корни полинома B относительно z_2. Для обеспечения устойчивости необходимо, чтобы все корни относительно z_2 были больше единицы при значении z_1 меньше единицы.

Пример 1 (неустойчивый фильтр)

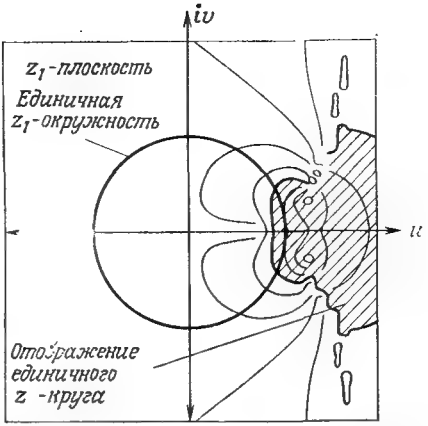
Рассмотрим фильтр с передаточной функцией $H(z_1, z_2) = 1/B(z_1, z_2)$, где

$$B(z_1, z_2) = 1 - 1{,}5z_1 + 0{,}6z_1^2 - 1{,}2z_2 + 1{,}8z_1z_2 - 0{,}72z_1^2z_2 + 0{,}5z_2^2 - 0{,}75z_1z_2^2 + 0{,}25z_1^2z_2^2.$$

Положив $z_1 = \hat{z}_1$, получаем

$$B(\hat{z}_1, z_2) = (1 - 1{,}5\hat{z}_1 + 0{,}6\hat{z}_1^2) + (-1{,}2 + 1{,}8\hat{z}_1 + 0{,}72\hat{z}_1^2) z_2 + (0{,}5 - 0{,}75\hat{z}_1 + 0{,}25\hat{z}_1^2) z_2^2. \quad (4.6)$$

Уравнение (4.6) — это квадратное уравнение относительно z_2 с комплексными коэффициентами. Таким образом, можно вычислить значения корней этого уравнения и графически отобразить их в виде функции координат точек размещения на z_1-плоскости. На фиг. 4.3 показана контурная диаграмма для одного из корней относительно z_2 в функции комплексной переменной



Фиг. 4.3. Отображение единичного z_2-круга и единичная z_1-окружность для неустойчивого фильтра.

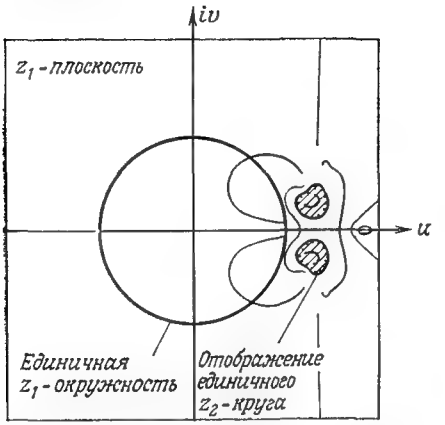
$z_1 = u + iv$. Выбранный корень относительно z_2 имеет наименьшее значение, поскольку требуется найти такие корни, для которых справедливо соотношение $|z_2| \leqslant 1$. Контурные линии построены с приращением уровня на 0,1; жирная контурная линия соответствует значению $z_2 = 1$. Заштрихованная область соответствует соотношению $|z_2| < 1$ и поэтому представляет собой отображение области, ограниченной единичной z_2-окружностью, в z_1-плоскость при $B(z_1, z_2) = 0$. Рассматриваемый фильтр неустойчив, поскольку заштрихованная область пересекается единичной z_1-окружностью.

Пример 2 (устойчивый фильтр)

На фиг. 4.4 показано отображение единичного z_2-круга в z_1-плоскость для полинома

$$B(z_1, z_2) = 1 - 1{,}2z_2 + 0{,}5z_2^2 - 1{,}5z_1 + 1{,}8z_1z_2 - 0{,}75z_1z_2^2 + 0{,}6z_1^2 - \\ - 0{,}72z_1^2z_2 + 0{,}29z_1^2z_2^2 = 0. \quad (4.7)$$

Как и прежде, заштрихованная область обозначает отображение единичного z_2-круга в z_1-плоскость. В данном случае отображенный z_2-круг не пересекается единичной z_1-окружностью, т. е. фильтр с передаточной функцией $H(z_1, z_2) = 1/B(z_1, z_2)$ устойчив. Фиг. 4.3 и 4.4 иллюстрируют тот факт, что корневые диаграммы симметричны относительно оси u. Благодаря этому можно ограничиться построением только верхней половины z_1-плоскости.



Фиг. 4.4. Отображение единичного z_2-круга и единичная z_1-окружность для устойчивого фильтра.

Условия устойчивости для фильтра первого порядка

В том особом случае, когда в знаменателе используется полином первого порядка по обоим направлениям, отображение z_1-плоскости в z_2-плоскость или отображение z_2-плоскости в z_1-плоскость является билинейным, что позволяет наложить особые ограничения на коэффициенты фильтра. Пусть $H(z_1, z_2) = 1/B(z_1, z_2)$, где $B(z_1, z_2) = 1 + b_{21}z_1 + b_{12}z_2 + b_{22}z_1z_2$. Фильтр с характеристикой $F(z_1, z_2)$ будет устойчивым, если и только если удовлетворяются следующие условия:

Условие 1

$$\left|\frac{1}{b_{12}}\right| > 1.$$

Условие 2

$$\left|\frac{1+b_{12}}{b_{21}+b_{22}}\right| > 1. \qquad (4.8)$$

Условие 3

$$\left|\frac{1-b_{12}}{b_{21}-b_{22}}\right| > 1.$$

Эти три условия были сформулированы Шэнксом в его неопубликованных выкладках (1969 г.). Доказательство было представлено Хуангом [1], который воспользовался теоремой 2 (см. ниже).

Примеры

Рассмотрим фильтр с передаточной функцией

$$H(z_1, z_2) = 1/(1 - 0{,}7z_1 - 0{,}5z_2 + 0{,}3z_1, z_2). \qquad (4.9)$$

Коэффициенты, используемые в (4.9), удовлетворяют всем условиям (4.8). Следовательно, фильтр с передаточной функцией (4.9) устойчив.

Пример неустойчивого фильтра дает передаточная функция вида

$$H(z_1, z_2) = 1/(1 - 0{,}95z_2 - 0{,}95z_1 + 0{,}5z_1, z_2). \qquad (4.10)$$

Коэффициенты фильтра удовлетворяют условиям 1 и 3 в (4.8), но не удовлетворяют условию 2.

Применение теоремы 1 для проверки устойчивости связано с нахождением корней полинома бесконечное число раз; на практике достаточно ограничиться нахождением корней в большом числе точек. Существует способ понизить «размерность» процедуры проверки на устойчивость, основанный на следующей теореме Хуанга.

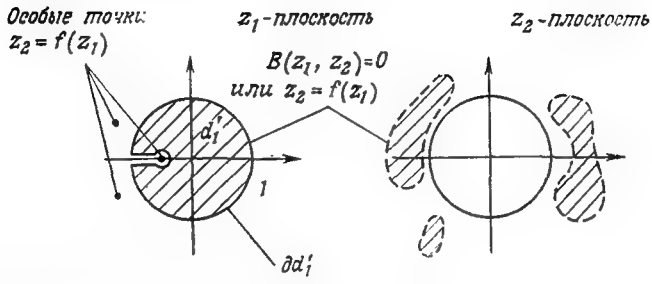
4.2.2. Теорема 2 (теорема Хуанга)

Физически реализуемый фильтр с z-преобразованием $H(z_1, z_2) = 1/B(z_1, z_2)$, где B — полином, устойчив, если и только если: 1) отображение $\partial d_1 \equiv (z_1;\ |z_1| = 1)$ в z_2-плоскость, соответствующее $B(z_1, z_2) = 0$, расположено вне области $d_2 \equiv \equiv (z_2;\ |z_2| \leqslant 1)$; 2) ни одна точка области $d_1 = (z_1;\ |z_1| \leqslant 1)$ не отображается в точку $z_2 = 0$ при выполнении условия $B(z_1, z_2) = 0$.

Доказательство. Требуется доказать, что условия устойчивости теорем 1 и 2 эквиваленты. Очевидно, что из условий устойчивости теоремы 1 вытекают условия устойчивости теоремы 2. Поэтому мы будем доказывать обратное соответствие.

Полином от двух переменных $B(z_1, z_2) = 0$ определяет алгебраическую функцию [2] $z_2 = f(z_1)$. Во-первых, следует видоизменить конфигурацию единичной окружности в z_1-плоскости, чтобы исключить попадание особых точек функции f в область, охватываемую этой окружностью. Модифицированная единичная окружность $\partial d_1'$ показана на фиг. 4.5; d_1' обозначает замкнутую область, ограниченную контуром $\partial d_1'$. Точка $z_1 = z_1^0$ называется особой точкой функции $z_2 = f(z_1)$, если уравнение $B(z_1^0, z_2) = 0$, рассматриваемое как уравнение относительно z_2, имеет множество (конечно- или бесконечномерное) корней.

Согласно теории алгебраических функций, функция $z_2 = \equiv f(z_1)$ имеет в области d_1' несколько ветвей, каждая из которых является аналитической. Следовательно, как вытекает из теоремы о максимуме модуля, максимум $|f(z_1)|$ в области d_1' располагается на контуре $\partial d_1'$, а минимум $|f(z_1)|$ в области d_1' может существовать внутри этой области только в том случае,



Фиг. 4.5. Отображение $z_2 = f(z_1)$ (по Хуангу).

если он имеет нулевое значение. Однако условие 2 теоремы 2 гласит, что функция $f(z_1)$ в области d_1' нигде не равна нулю. Следовательно, минимум $f(z_1)$ расположен на контуре $\partial d_1'$, т. е.

$$|f(d_1')| \geqslant \min |f(\partial d_1')|, \tag{4.11}$$

откуда следует, что если $|f(\partial d_1')| > 1$, то $|f(d_1')| > 1$; таким образом, чтобы гарантировать расположение $f(d_1')$ вне единичного круга $d_2 = (z_2, |z_2| \leqslant 1)$, достаточно гарантировать расположение $f(\partial d_1)$ вне области d_2.

Мы уже почти достигли цели, но окончательный результат еще не получен. В действительности требуется доказать, что если $|f(\partial d_1)| > 1$, то $|f(d_1)| > 1$, где $d_1 \equiv (z_1, |z_1| \leqslant 1)$ и $\partial d_1 \equiv \equiv (z_1; |z_1| = 1)$. Поскольку обход по контуру $\partial d_1'$ может выполняться по любому пути, ведущему от ∂d_1 к особой точке, остается только доказать, что $|f(s)| > 1$, где s — особая точка. Поскольку каждая ветвь $z_2 = f(z_1)$ непрерывна в точке $z_1 = s$ и поскольку $|f(s + \varepsilon e^{i\theta})| > 1$ при произвольно малом ε и при любом θ, имеем $|f(s)| > 1$ (что и требовалось доказать).

Условия теоремы 2 сформулированы для того случая, когда единичная окружность z_1-плоскости отображается в z_2-плоскость. Однако теорему можно сформулировать и для случая отображения единичной окружности z_2-плоскости в z_1-плоскость. Эта формулировка представлена как следствие 2.

Следствие 2. Физически реализуемый фильтр с z-преобразованием $H(z_1, z_2) = 1/B(z_1, z_2)$ (где B — полином) устойчив, если и только если: 1) отображение $\partial d_2 \equiv (z_2; |z_2| = 1)$ в z_1-плоскость, соответствующее условию $B(z_1, z_2) = 0$, расположено вне области $d_1 \equiv (z_1; |z_1| \leqslant 1)$; 2) ни одна точка области $d_2 = = (z_2; |z_2| \leqslant 1)$ не отображается в точку $z_1 = 0$ при выполнении условия $B(z_1, z_2) = 0$.

Это следствие можно проиллюстрировать с помощью корневой диаграммы на фиг. 4.4. Здесь отображения единичной

z_2-окружности, очевидно, представляют собой непрерывные контурные линии, построенные по точкам расположения корней $B(z_1, |z_2| = 1) = 0$. В данном примере условие 1 следствия 2 выполняется. Условие 2 следствия 2 также выполняется, поскольку внутренняя область единичного z_2-круга отображается в область, ограниченную отображением единичного z_2-круга. Оба условия удовлетворяются, поэтому фильтр устойчив. Условие 1 следствия 2 гарантирует, что отображение единичной z_2-окружности не пересекает единичного z_1-круга. Условие 2 дополнительно гарантирует, что отображение единичного z_2-круга не пересекает ни одной точки единичного z_1-круга.

На фиг. 4.3 условие 2 не выполняется. Отображение единичной z_2-окружности пересекает единичную z_1-окружность, поэтому фильтр является неустойчивым.

Методика проверки на устойчивость

Чтобы произвести проверку на устойчивость с помощью теоремы 2, необходимо отобразить $\partial d_1 = (z_1; |z_1| = 1)$ в z_2-плоскость согласно $B(z_1, z_2) = 0$ и проверить, расположено ли отображение вне $d_2 \equiv (z_2; |z_2| \leqslant 1)$. Кроме того, требуется решить уравнение $B(z_1, 0) = 0$, чтобы проверить, имеется ли хотя бы один корень, значение которого меньше единицы.

Известны два метода применения рассмотренной теоремы как основы процедуры проверки на устойчивость, дающие результат через конечное число шагов. Первый метод основан на использовании проверки Гурвица; описание этого метода дал Хуанг [1]. Второй метод, основанный на использовании матрицы Шура — Кона, был предложен Андерсоном и Джури [3].

Метод Хуанга с использованием проверки Гурвица. Согласно Андерсону и Джури, условия теоремы 2 можно сформулировать следующим образом:

$B(z_1, z_2) \neq 0$ при $|z_1| \leqslant 1 \cap |z_2| \leqslant 1$, если и только если удовлетворяются два условия:

$$B(z_1, 0) \neq 0, \quad |z_1| \leqslant 1, \tag{4.12}$$

$$B(z_1, z_2) \neq 0, \quad |z_1| = 1, \quad |z_2| \leqslant 1. \tag{4.13}$$

Проверка условия (4.12) не вызывает трудностей, поскольку $B(z_1, 0)$ — полином от одной переменной, и имеется несколько методов проверки того, все ли его нули расположены вне единичной окружности. Одна группа проверочных процедур основана на замене $B(z_1, 0)$ другим полиномом, получаемым билинейным преобразованием, и проверке последнего на принадлежность к полиномам Гурвица. (Некоторый полином является полиномом Гурвица, если все его нулевые корни имеют отрица-

тельные действительные части.) Методов же выявления полиномов Гурвица очень много (см., например, [4]).

Метод Андерсона и Джури с использованием матрицы Шура — Кона. Второй метод основан на использовании матрицы Шура — Кона [5, 6, 8]. Это квадратная эрмитова матрица, размер которой равен степени полинома $B(z_1, 0)$; элементы матрицы являются простыми функциями коэффициентов $B(z_1, 0)$. Матрица является отрицательно определенной, если и только если все нули $B(z_1, 0)$ расположены в области $|z_1| > 1$. Установить тот факт, что матрица является отрицательно определенной, можно, очевидно, путем проверки знаков ее ведущего главного минора.

Некоторые другие методы проверки можно найти в работах [7, 11, 12]. Среди них имеется один метод, обладающий высокой вычислительной эффективностью, который основан на построении рекурсивным способом конечномерного множества полиномов, каждый элемент которого имеет более низкую степень, чем предыдущий, причем $B(z_1, 0)$ является первым элементом этого множества. Рассмотрение некоторых коэффициентов таких полиномов позволяет быстро определить, удовлетворяется условие (4.12) или нет.

Проверка выполнения условия (4.13) более сложна, чем в случае условия (4.12), тем не менее такую проверку можно произвести за конечное число шагов. Как и при проверке условия (4.12), в данном случае можно применить два метода, один из которых основан на использовании результатов работы Ансела и описан Хуангом [1], а другой, представленный Андерсоном и Джури [3], — на использовании матрицы Шура — Кона. Оба метода требуют длительных выкладок. Заинтересованный читатель может обратиться к упомянутым источникам.

4.3 Проектирование

Как отмечалось выше, проектирование рекурсивных фильтров требует решения двух основных проблем: проблемы аппроксимации и проблемы устойчивости. Основные теоретические сведения о методах проверки устойчивости двумерных рекурсивных фильтров представлены в предыдущем разделе. Теперь будут рассмотрены проблемы выбора коэффициентов рекурсивного фильтра и методы обеспечения его устойчивости.

Существует много способов выбора коэффициентов двумерного рекурсивного фильтра. Здесь мы рассмотрим три способа.

1. Обобщение одномерных методов.

2. Выбор коэффициентов фильтра таким образом, чтобы его импульсная характеристика аппроксимировала заданную импульсную характеристику.

3. Выбор коэффициентов фильтра с использованием методов аппроксимации таким образом, чтобы его частотная характеристика аппроксимировала заданную частотную характеристику.

4.3.1. Обобщение одномерных методов

В одномерном случае можно привести множество методов проектирования рекурсивных фильтров в частотной области [13, 14]. Большая часть работ в области одномерного синтеза в частотной области посвящена проектированию фильтров нижних частот и полосовых фильтров. Некоторые наиболее эффективные методы проектирования двумерных рекурсивных фильтров основаны на обобщении одномерных методов.

В одномерном случае в качестве расчетных параметров часто используют частоты среза, т. е. частоты точек перехода между «полосой пропускания» и «полосой задерживания». В двумерном случае имеется переходная область, разделяющая области пропускания и задерживания. Конфигурация переходной области определяется требованиями проектировщика к фильтру и числом коэффициентов, которые желательно иметь в фильтре.

Разделимые фильтры

Простейший случай представляет собой разделимый фильтр, т. е. фильтр с передаточной функцией вида

$$F(z_1, z_2) = F_1(z_1) F_2(z_2). \tag{4.14}$$

Таким образом, двумерный фильтр составляется из двух одномерных фильтров, соединенных последовательно. С разделимым фильтром связана четырехквадрантная симметрия в частотной области, которая иногда оказывается нежелательной.

Для фильтров с действительными коэффициентами характерна симметрия относительно прямой, проходящей через начало координат двумерной частотной плоскости. Ниже рассматриваются два метода проектирования разделимых фильтров с действительными коэффициентами, причем эти методы позволяют избежать четырехквадрантной (в двумерной частотной плоскости) симметрии, связанной с (4.14).

Поворот частотной характеристики одномерного фильтра

Предположим, что спроектирован фильтр нижних частот или полосовой фильтр, имеющий одно измерение. Такой фильтр можно рассматривать как двумерный фильтр, свойства которого изменяются только в одном направлении. Если воспользоваться переменными (s_1, s_2) двумерного преобразования Лапласа [15], то такой фильтр можно определить как

$$F(s_1, s_2) = F(s_2). \tag{4.15}$$

Частотную характеристику этого фильтра можно определить, положив $s_1 = i2\pi f_1$ и $s_2 = i2\pi f_2$. Очевидно, что частотная характеристика изменяется только в направлении частотной оси f_2. Предположим теперь, что требуется выполнить преобразование, или отображение (s_1, s_2)-плоскости в (s_1', s_2')-плоскость, такое, что полоса пропускания $F(s_1, s_2)$ отобразится в подходящую область (s_1', s_2')-плоскости. В частности, рассмотрим вращение осей (s_1, s_2)-плоскости.

Метод преобразования. Повернуть (s_1, s_2)-оси на угол β можно с помощью преобразования

$$s_1 = s_1' \cos\beta + s_2' \sin\beta, \tag{4.16}$$

$$s_2 = s_2' \cos\beta - s_1' \sin\beta. \tag{4.17}$$

Поворот (s_1, s_2)-осей также вызывает поворот осей двумерной частотной (f_1, f_2)-плоскости. Рассмотрим фильтр с передаточной функцией $F(s_1, s_2)$. Используя (4.16) и (4.17), можно повернуть (f_1, f_2)-оси на угол β, чтобы они заняли новое положение (f_1', f_2'). Однако эта операция вызывает поворот области «пропускания» на угол $-\beta$ относительно новых частотных осей.

В качестве примера рассмотрим фильтр с передаточной функцией вида

$$F(s_1, s_2) = \frac{1}{s_2 + \alpha}. \tag{4.18}$$

Амплитудно-частотная характеристика (АЧХ) этого фильтра изменяется только по оси f_2. Выполнив преобразование согласно (4.16) и (4.17), получаем

$$F'(s_1', s_2') = \frac{1}{s_2' \cos\beta - s_1' \sin\beta + \alpha}. \tag{4.19}$$

АЧХ нового фильтра изменяется по обеим частотным осям.

Можно спроектировать двумерные полосовые фильтры, частотные характеристики которых ориентированы под определенным углом относительно (f_1, f_2)-осей. Прежде всего, следует спроектировать подходящий одномерный фильтр нижних частот или полосовой фильтр, используя одномерную s-область. Выразив уравнение этого фильтра через s_1 или s_2, произведем поворот его осей на требуемый угол в (s_1', s_2')-плоскости с помощью (4.16) или (4.17). Результирующее выражение $F'(s_1', s_2')$ описывает непрерывный двумерный фильтр.

Остается спроектировать эквивалентный двумерный дискретный фильтр. Для этого можно воспользоваться билинейным

z-преобразованием [14]. Таким образом, имеем

$$s_1' = \frac{2(1-z_1)}{\Delta_1(1+z_1)}, \tag{4.20}$$

$$s_2' = \frac{2(1-z_2)}{\Delta_2(1+z_2)}, \tag{4.21}$$

где Δ_1 и Δ_2 — периоды дискретизации. Применяя к $F'(s_1', s_2')$ уравнения билинейного преобразования, согласно (4.19), получаем

$$F(z_1, z_2) = \frac{\Delta_1\Delta_2}{2} \frac{(1+z_1)(1+z_2)}{(1+b_{21}z_1+b_{12}z_2+b_{22}z_1z_2)}, \tag{4.22}$$

где

$$b_{21} = \left(\frac{\alpha\Delta_1\Delta_2}{2} + \Delta_1\cos\beta + \Delta_2\sin\beta\right)/D, \tag{4.23}$$

$$b_{12} = \left(\frac{\alpha\Delta_1\Delta_2}{2} - \Delta_1\cos\beta - \Delta_2\sin\beta\right)/D, \tag{4.24}$$

$$b_{22} = \left(\frac{\alpha\Delta_1\Delta_2}{2} - \Delta_1\cos\beta + \Delta_2\sin\beta\right)/D \tag{4.25}$$

и

$$D = \frac{\alpha\Delta_1\Delta_2}{2} + \Delta_1\cos\beta - \Delta_2\sin\beta.$$

Эти выражения можно упростить, положив $\Delta_1 = \Delta_2 = \Delta$. В (4.18) константа α выражает частоту среза (по уровню — 3 дБ), измеряемую в радиан/единица [16][1]). Пусть

$$\alpha = 2\pi f_c f_N,$$

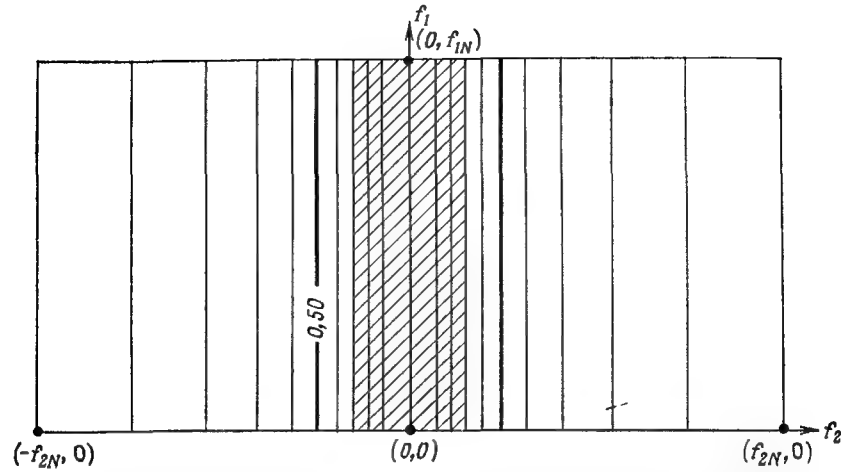
где f_c — частота среза, выражаемая в долях частоты Найквиста; f_N — частота Найквиста, измеряемая числом периодов, приходящихся на единицу измерения. Поскольку $f_N = {}^1/_2\Delta$,

$$\alpha = \frac{\pi f_c}{\Delta}.$$

Таким образом, получаем

$$b_{21} = [\pi f_c + 2(\cos\beta + \sin\beta)]/D,$$
$$b_{12} = [\pi f_c - 2(\cos\beta + \sin\beta)]/D,$$
$$b_{22} = [\pi f_c - 2(\cos\beta - \sin\beta)]/D,$$
$$D = [\pi f_c + 2(\cos\beta - \sin\beta)].$$

[1]) Единица измерения определяется решаемой задачей. В случае регистрации сейсмических сигналов используют футы в одном направлении и выбранные единицы измерения времени — в другом. В случае фотографической регистрации по двум координатам используются пространственные единицы измерения.



Фиг. 4.6. Двумерная АЧХ, соответствующая передаточной функции $(1 + z_2)/(1 - 0{,}6186z_2)$ (f_{1N} и f_{2N} — частоты Найквиста по осям f_1 и f_2 соответственно).

Примеры. Рассмотрим фильтр с $F(s) = 1/(s + \alpha)$, где α выбрано таким, что частота среза составляет 0,15 частоты Найквиста. Таким образом, $f_c = 0{,}15$ и для $\beta = 0°$ имеем

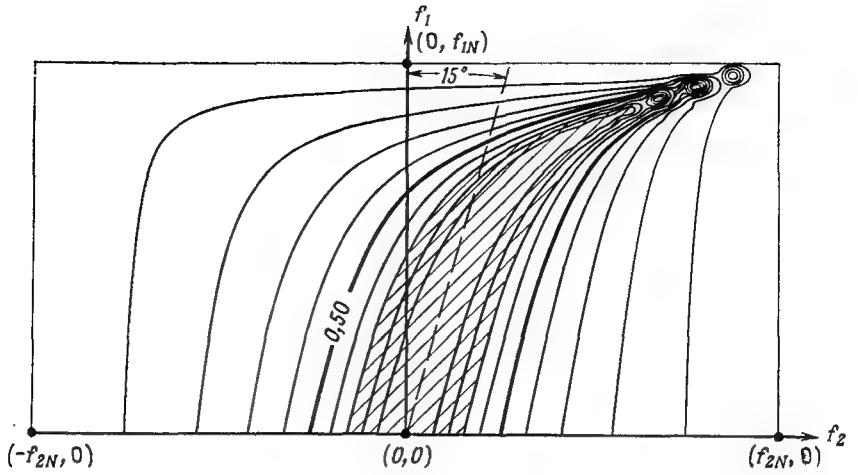
$$F(z_1, z_2) = \frac{\Delta^2}{2} \frac{(1 + z_1)(1 + z_2)}{1 - 0{,}6186z_2 + z_1 - 0{,}6186z_1z_2} = \\ = \frac{\Delta^2}{2} \frac{1 + z_2}{1 - 0{,}6186z_1}.$$

На фиг. 4.6 представлена контурная диаграмма АЧХ рассматриваемого фильтра, нормированной относительно максимального коэффициента передачи 1. Контурные линии построены с шагом 0,1. В пределах области пропускания, отмеченной штриховкой, значение АЧХ превышает 0,7. Поскольку $\beta = 0$, АЧХ не изменяется по оси f_1. По оси f_2 АЧХ соответствует фильтру нижних частот. Следует обратить внимание на то, что частота среза составляет 15% частоты Найквиста; это согласуется с расчетным значением $f_c = 0{,}15$.

Для $\beta = 15°$ получаем

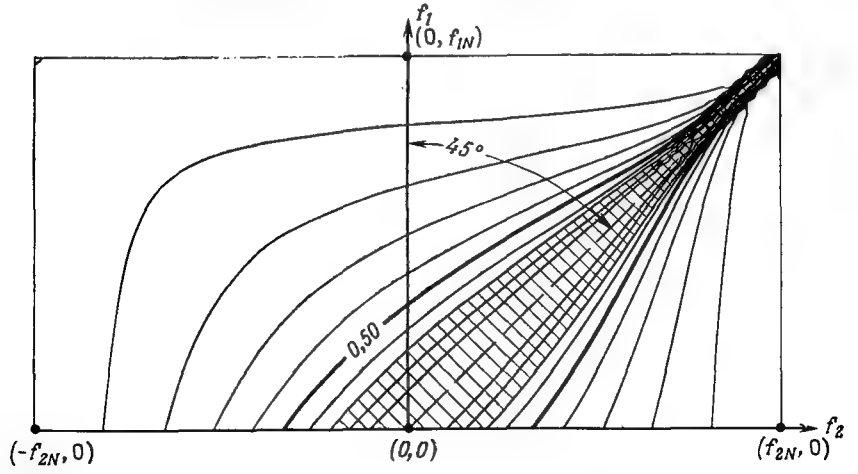
$$F(z_1, z_2) = \frac{\Delta^2}{2} \frac{(1 + z_1)(1 + z_2)}{1 - 0{,}3229z_2 + 0{,}6455z_1 - 0{,}6773z_1z_2}.$$

На фиг. 4.7 представлена контурная диаграмма АЧХ нового фильтра. Контурные линии также построены с шагом 0,1. Как и требовалось, АЧХ повернута на угол 15° относительно оси f_2.

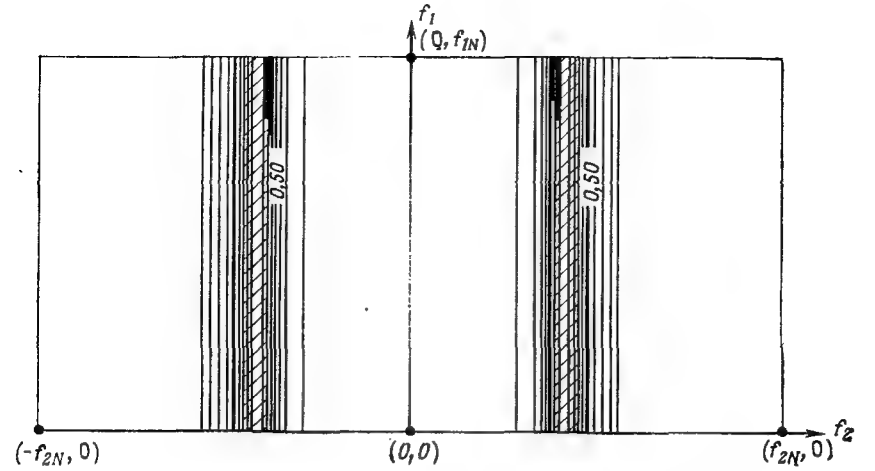


Фиг. 4.7. АЧХ фильтра нижних частот после поворота на угол 15°.

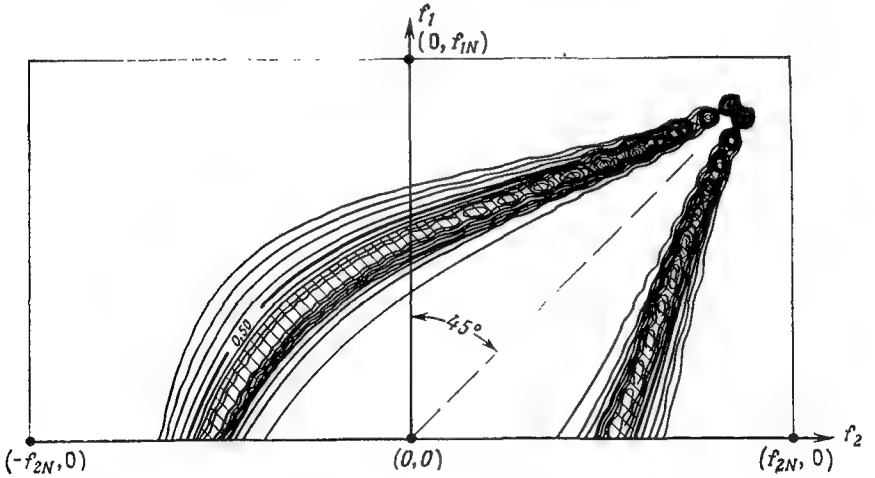
Однако теперь контурные линии уже не являются прямыми, причем заметна тенденция к возрастанию искажений по мере увеличения частот (f_1, f_2). Этот эффект обусловлен применением билинейного (z_1, z_2)-преобразования. В результате подстановки (4.20) и (4.21) в выражение (4.19) в числителе выражения для $F(z_1, z_2)$ появляются множители $(1 + z_1)(1 + z_2)$. Первый из них обеспечивает нулевое значение $F(z_1, z_2)$ на всех



Фиг. 4.8. АЧХ фильтра нижних частот после поворота на угол 45°.



Фиг. 4.9. АЧХ полосового фильтра.

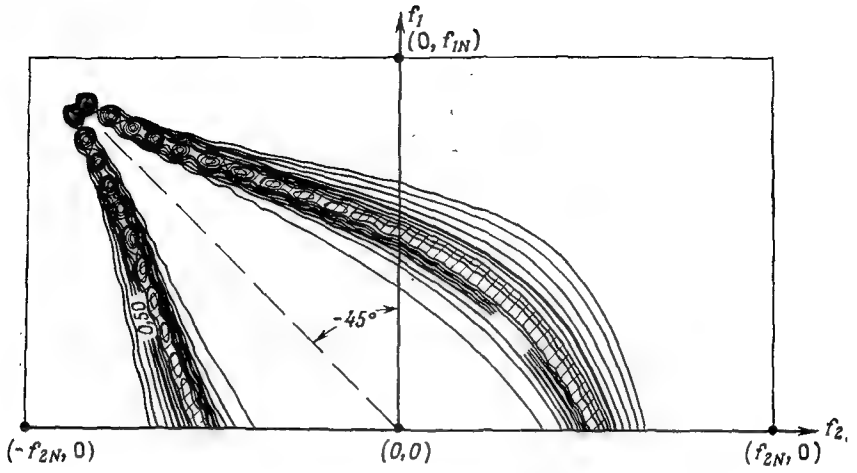


Фиг. 4.10. АЧХ полосового фильтра после поворота на угол 45°.

частотах, соответствующих равенству f_1 частоте Найквиста. Соответственно множитель $(1 + z_2)$ обеспечивает нулевое значение $F(z_1, z_2)$ при равенстве f_2 частоте Найквиста.

На фиг. 4.8 показана АЧХ фильтра, повернутая на 45°. Передаточная функция фильтра имеет вид

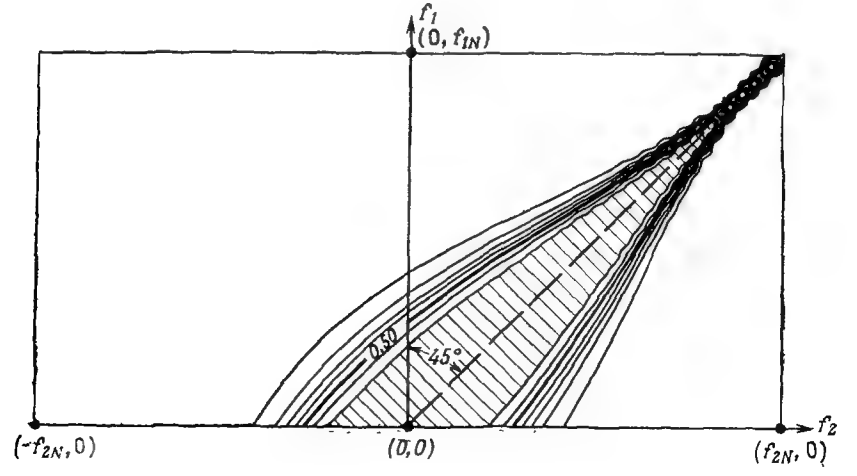
$$F(z_1, z_2) = \frac{(1 + z_1)(1 + z_2)}{1 + 0{,}1428z_2 + 0{,}1428z_1 - 0{,}7144z_1z_2}.$$



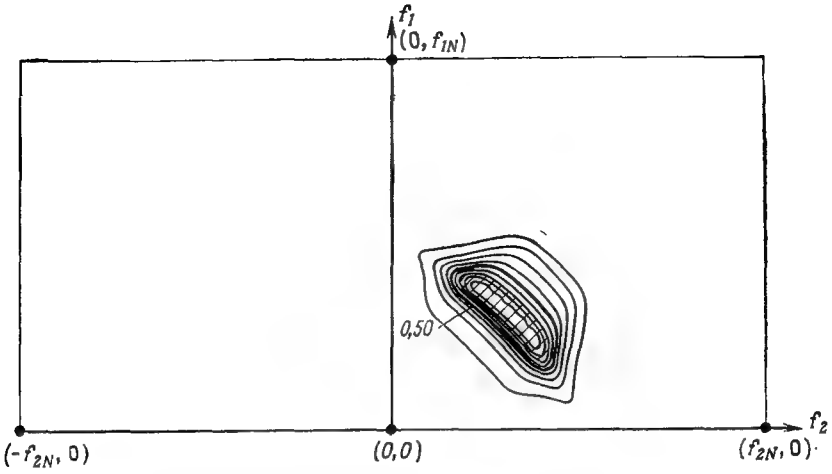
Фиг. 4.11. АЧХ полосового фильтра после поворота на угол —45°.

Можно также «повернуть» АЧХ полосового фильтра. На фиг. 4.9 показана АЧХ полосового фильтра, которая получена путем билинейного преобразования АЧХ фильтра Баттерворта восьмого порядка [17]. Фиг. 4.10 иллюстрирует АЧХ того же фильтра, повернутую на 45°, причем и в этом случае применено билинейное преобразование.

Поворот на отрицательный угол можно обеспечить, видоизменив методику фильтрации данных. Пусть $D(z_1, z_2)$ — входной



Фиг. 4.12. АЧХ фильтра нижних частот после поворота на угол 45°.



Фиг. 4.13. АЧХ фильтра, полученного путем последовательного соединения двух фильтров с АЧХ, приведенными на фиг. 4.11 и 4.12.

массив размера $M \times N$, а $F(z_1, z_2)$ описывает фильтр с АЧХ, показанной на фиг. 4.10. Мы не будем выполнять фильтрацию $F(z_1, z_2)$ обычным способом, как описывает выражение (4.3). Напротив, начнем фильтрацию с входной точки $d_{1,N}$ и продолжим ее в обратном направлении оси y. Такая процедура эквивалентна применению фильтра с передаточной функцией вида

$$G(z_1, z_2) = F(z_1, 1/z_2).$$

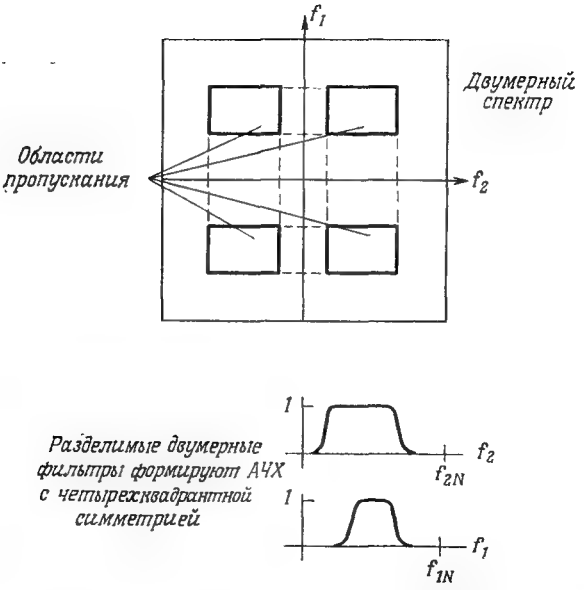
Новый фильтр имеет АЧХ, показанную на фиг. 4.11; как видно, полоса пропускания повернута на угол —45°.

Рассмотренные «повернутые» фильтры можно использовать совместно для получения других конфигураций АЧХ. Пусть имеется фильтр с АЧХ, показанной на фиг. 4.12. Эта характеристика была получена путем поворота на угол 45° АЧХ фильтра нижних частот. Предположим, что данные фильтруются указанным фильтром и фильтром с АЧХ на фиг. 4.11. Результирующая АЧХ будет представлять собой произведение этих двух АЧХ. Последовательно соединенные фильтры имеют АЧХ, показанную на фиг. 4.13.

Сдвиг частотной характеристики двумерного фильтра нижних частот [1])

Другой исключительно полезный метод проектирования двумерных рекурсивных фильтров состоит в использовании сдвига частотных характеристик. Если двумерный рекурсивный фильтр

[1]) Этот материал принадлежит одному из авторов (Р. Риду).

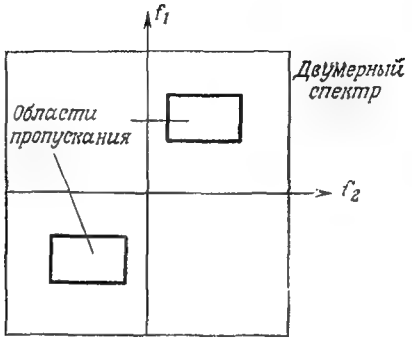


Фиг. 4.14. Четырехквадрантная АЧХ двумерного полосового фильтра, полученного путем последовательного соединения двух полосовых одномерных фильтров.

является действительным (имеет действительные коэффициенты) и разделимым, то его АЧХ обладает четырехквадрантной симметрией в двумерной частотной области. Фильтры такого рода можно описать с помощью выражения (4.14). Четырехквадрантную симметрию АЧХ полосового фильтра иллюстрирует фиг. 4.14.

Предположим теперь, что одномерные передаточные функции в (4.14) имеют комплексные коэффициенты. В этом случае АЧХ одномерных фильтров с передаточными функциями $F_1(z_1)$ и $F_2(z_2)$ уже не будут обладать знакомой нам симметрией относительно начала координат частотной области. Последовательное соединение таких двух фильтров позволяет создать разделимый двумерный фильтр, однако АЧХ этого фильтра не будет обладать какой-то определенной симметрией в двумерной частотной области. Следует заметить, что фильтр с комплексными коэффициентами не часто оказывается полезным при обработке действительных данных. Объединяя действительные части передаточных функций двух комплексных одномерных фильтров, включенных последовательно, мы получаем

$$\begin{aligned}\operatorname{Re}\{F_1\}\operatorname{Re}\{F_2\} &= \frac{1}{2}(F_1 + F_1^*)\frac{1}{2}(F_2 + F_2^*) = \\ &= (F_1F_2 + F_1F_2^* + F_2F_1^* + F_1^*F_2^*)/4, \qquad (4.26)\end{aligned}$$



Фиг. 4.15 АЧХ неразделимого двумерного полосового фильтра.

где Re $\{\ldots\}$ указывает на использование только действительных частей коэффициентов фильтров, а звездочка (*) обозначает передаточную характеристику исходного фильтра с комплексно сопряженными коэффициентами. Фильтр, описываемый (4.26), является действительным и разделимым.

Если перемножить передаточные функции комплексных одномерных фильтров, включенных последовательно, и выделить действительную часть результата, то получим

$$\mathrm{Re}\{F_1F_2\} = \frac{1}{2}(F_1F_2 + F_1^*F_2^*). \tag{4.27}$$

Результирующий фильтр имеет действительные коэффициенты, не является разделимым и обладает требуемой симметрией по прямой, проходящей через начало координат двумерной частотной плоскости.

Отсюда следует, что всегда можно создать фильтр с действительными коэффициентами, АЧХ которого имеет области пропускания только в двух противолежащих квадрантах (фиг. 4.15). Ниже описан метод осуществления такого преобразования.

Метод преобразования. Здесь будет показано, каким образом сформировать фильтр с двуквадрантной АЧХ, используя два одномерных рекурсивных фильтра с передаточными функциями вида

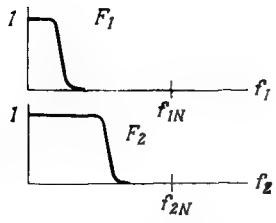
$$F_1(z_1) = \frac{A_1(z_1)}{B_1(z_1)}, \qquad F_2(z_2) = \frac{A_2(z_2)}{B_2(z_2)}, \tag{4.28}$$

где A_1 и B_1 — полиномы:

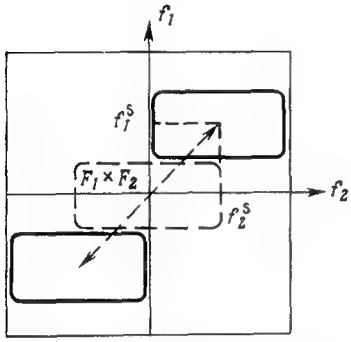
$$A_1(z_1) = \sum_{i=1}^{N_1} a_i z_1^{i-1},$$

$$B_1(z_1) = \sum_{j=1}^{M_1} b_j z_1^{j-1}.$$

Полиномы A_2 и B_2 определяются аналогично.



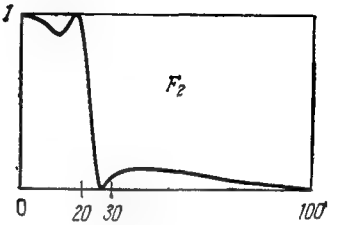
Перемножение $F_1(z_1) \times F_2(z_2)$ дает:

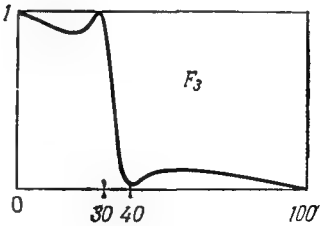


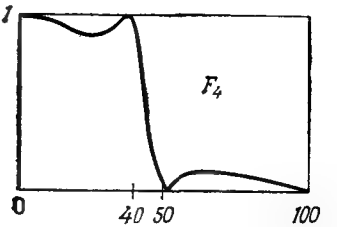
Сдвиг АЧХ $F_1 \times F_2$ фильтра нижних частот для получения двухквадрантной АЧХ

Фиг. 4.16. Пример формирования АЧХ неразделимого полосового фильтра путем сдвига АЧХ, полученной в результате последовательного соединения двух одномерных фильтров нижних частот (f_{1N} и f_{2N} — частоты Найквиста по осям f_1 и f_2 соответственно).

Метод формирования фильтра с двуквадрантной АЧХ предполагает наличие двух действительных одномерных фильтров нижних частот с передаточными функциями $F_1(z_1)$ и $F_2(z_2)$. Следующий шаг — сдвиг этих двух передаточных функций вдоль осей f_1 и f_2 одномерных частотных областей таким образом, чтобы они заняли требуемое положение. Полученные при этом фильтры будут иметь комплексные коэффициенты, однако их можно объединить в соответствии с (4.27) и получить полосовой фильтр с требуемой двуквадрантной АЧХ. Следует отметить, что выделение действительных частей коэффициентов означает, что результирующий фильтр в общем случае уже не является разделимым. Таким образом, если известны передаточная функция F_1 одномерного фильтра нижних частот, передаточная функция F_2 аналогичного фильтра, угол θ_1, соответствующий частотному сдвигу (или повороту в z_1-плоскости) F_1, и угол θ_2, соответствующий частотному сдвигу F_2, то результирующую передаточную функцию можно сделать расщепленной (как показано на фиг. 4.16) с помощью следующих преобразований.







Фиг. 4.17. АЧХ одномерных фильтров нижних частот, которые были использованы для получения различных двумерных полосовых фильтров.

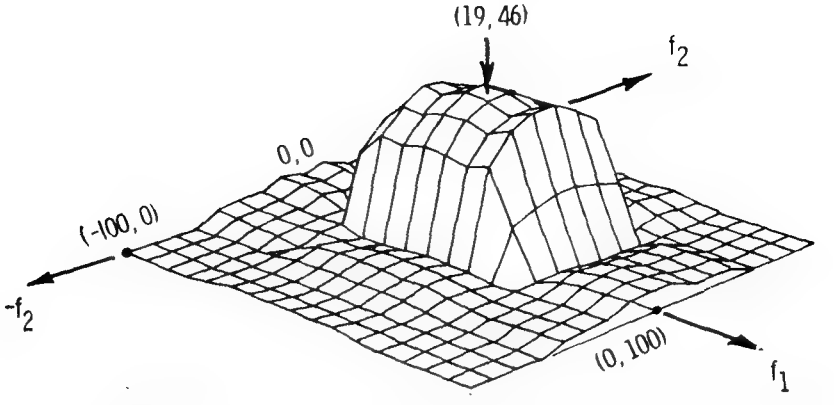
Произвести подстановку

$$z_1 = z_1' e^{-j\theta_1},$$
$$z_2 = z_2' e^{-j\theta_2}$$

в F_1 и F_2 соответственно. Это дает комплексные коэффициенты $\hat{A}_1$, $\hat{B}_1$, $\hat{A}_2$ и $\hat{B}_2$ и вызывает сдвиг соответствующих им одномерных частотных характеристик в двух частотных областях. Обозначая передаточные функции преобразованных фильтров $\hat{F}_1$ и $\hat{F}_2$, получаем

$$\hat{F}_1\hat{F}_2 + \hat{F}_1^*\hat{F}_2^* = \frac{2\,\mathrm{Re}\,\{\hat{A}_1\hat{B}_1\hat{A}_2\hat{B}_2\}}{|\hat{B}_1|^2\,|\hat{B}_2|^2}. \qquad (4.29)$$

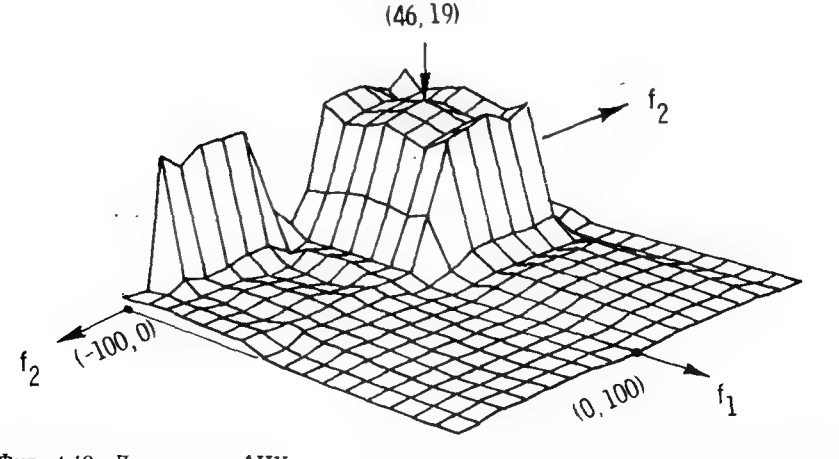
Передаточная функция результирующего фильтра имеет разделимый знаменатель с четными степенями z_1 и z_2. Если исходные



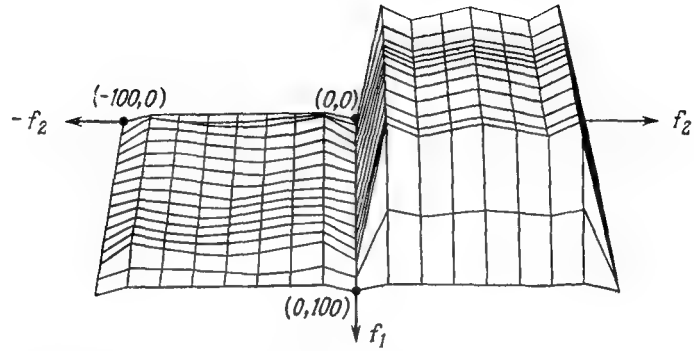
Фиг. 4.18. Двумерная АЧХ, полученная путем сдвига результата перемножения F_2 и F_3 в точку (19, 46).

фильтры с передаточными функциями F_1 и F_2 являются устойчивыми, то знаменатель подобного рода гарантирует устойчивость результирующего фильтра, поскольку нули полиномов B получены путем поворота нулей полиномов B.

Примеры. Уравнение (4.29) легко реализуется с помощью ЦВМ. В качестве примера применения описанного метода преобразования рассмотрим три эллиптических фильтра третьего порядка с пульсацией 0,1 в полосах пропускания и задерживания. Ширина переходной полосы каждого фильтра составляет 10 Гц при частоте Найквиста 100 Гц. Фильтры с передаточными



Фиг. 4.19. Двумерная АЧХ, полученная путем сдвига результата перемножения F_2 и F_3 в точку (46, 19).



Фиг. 4.20. Двумерная АЧХ, полученная путем сдвига результата перемножения F_4 и F_4 в точку (50, 50).

функциями F_2, F_3 и F_4 имеют полосы пропускания от 0 до 20, 30 и 40 Гц соответственно. АЧХ этих трех фильтров показаны на фиг. 4.17.

Результирующие двумерные АЧХ, полученные методом частотного сдвига, показаны на фиг. 4.18—4.20.

Как можно видеть на фиг. 4.18—4.20, описанный метод особенно полезен в случае формирования двумерных рекурсивных фильтров с полосами пропускания или задерживания прямоугольной конфигурации, размещенными в различных областях двумерной частотной плоскости. Кроме того, (4.29) с очевидностью показывает, что коэффициенты знаменателя образуют разделимый полином, а это означает уменьшение числа вычислительных этапов при нахождении значения каждой выходной точки. Из уравнения (4.29) следует, что АЧХ в каждом отдельном квадранте не является совершенно независимой от АЧХ в других квадрантах. АЧХ вне полосы пропускания в квадранте 1 (связанная с первым членом левой части уравнения) влияет на полосу пропускания в симметрично расположенной части квадранта 3. Однако это влияние обычно незначительно, если значения АЧХ фильтра пренебрежимо малы вне областей пропускания. Следует отметить, что значения АЧХ в каждом из квадрантов суммируются, что следует из (4.29). Если области пропускания сдвинутых АЧХ случайно наложатся, то вместо единичного уровня в результирующей области пропускания будет получен удвоенный уровень.

4.3.2. Синтез в пространственной области

Предположим, что нам известна двумерная импульсная характеристика дискретного фильтра, который будет использоваться для обработки входного массива данных. Возможность

применения фильтра обеспечит двумерная свертка или двумерное преобразование Фурье. Однако можно также аппроксимировать известную импульсную характеристику синтезированной импульсной характеристикой двумерного рекурсивного фильтра. В некоторых случаях применение рекурсивного фильтра оказывается более эффективным методом, чем вычисление свертки или выполнение фильтрации в частотной области.

Метод синтеза

Представленный здесь метод является обобщением одномерного синтеза [17—19]. Пусть d_{ij} — требуемая импульсная характеристика, где $i = 1, 2, 3, \ldots, M_d$ и $j = 1, 2, 3, \ldots, N_d$. При этом

$$D(z_1, z_2) = \sum_{i=1}^{M_d} \sum_{j=1}^{N_d} d_{ij} z_1^{i-1} z_2^{j-1}. \tag{4.30}$$

Пусть аппроксимирующий двумерный рекурсивный фильтр имеет передаточную функцию вида

$$F(z_1, z_2) = \frac{A(z_1, z_2)}{B(z_1, z_2)}, \tag{4.31}$$

где

$$A(z_1, z_2) = \sum_{i=1}^{M_a} \sum_{j=1}^{N_a} a_{ij} z_1^{i-1} z_2^{j-1},$$

$$B(z_1, z_2) = \sum_{i=1}^{M_b} \sum_{j=1}^{N_b} b_{ij} z_1^{i-1} z_2^{j-1},$$

$b_{11} = 1$,

M_a, N_a, M_b, N_b — произвольные (но фиксированные) параметры.

Из (4.31) следует, что $F(z_1, z_2) B(z_1, z_2) = A(z_1, z_2)$. Перемножение полиномов от z эквивалентно свертке массивов, поэтому

$$a_{mn} = \sum_{i=1}^{M_b} \sum_{j=1}^{N_b} b_{ij} f_{m-i+1,\, n-j+1}. \tag{4.32}$$

Выражение (4.32) определяет коэффициенты a_{mn} в области целых чисел $m = 1, 2, \ldots, M_a$, $n = 1, 2, \ldots, N_a$. Вне этой области коэффициенты a_{mn} равны нулю. Определим множество пар целых чисел как

$$I_a = \{(i, j) : 1 \leqslant i \leqslant M_a, \quad 1 \leqslant j \leqslant N_a\}.$$

Кроме того, определим $\hat{I}_a$ как множество всех других значений (m, n), превышающих нулевое значение:

$$\hat{I}_a = \{(m, n) : m > 0, \quad n > 0, \quad (m, n) \notin I_a\}.$$

Таким образом, должны быть определены значения a_{mn} при $(m, n) \in I_a$. При $(m, n) \in \hat{I}_a$ имеем $a_{mn} \equiv 0$. С учетом (4.32) и того факта, что $b_{11} = 1$, получаем

$$f_{mn} = -\sum_{\substack{i=1 \\ i \cdot j \neq 1}}^{M_b} \sum_{j=1}^{N_b} b_{ij} f_{m-i+1,\, n-j+1} \quad \text{при } (m, n) \in \hat{I}_a.$$

Если правильно выбрать b_{ij}, f_{mn} будет аппроксимировать требуемую импульсную характеристику d_{mn}. Таким образом, можно записать

$$d_{mn} \approx -\sum_{\substack{i=1 \\ i \cdot j \neq 1}}^{M_b} \sum_{j=1}^{N_b} b_{ij} d_{m-i+1,\, n-j+1} \quad \text{при} \quad (m, n) \in \{\hat{I}_a \cap I_d\}, \quad (4.33)$$

где значения (m, n) должны принадлежать пересечению множества $\hat{I}_a$ и множества

$$I_d = \{(m, n) : 1 \leqslant m \leqslant M_d, \quad 1 \leqslant n \leqslant N_d\}.$$

Пусть e_{mn} — ошибка, которую можно добавить к правой части (4.33) для получения равенства

$$d_{mn} = e_{mn} - \sum_{\substack{i=1 \\ i \cdot j \neq 1}}^{M_b} \sum_{j=1}^{N_b} b_{ij} d_{m-i+1,\, n-j+1} \quad \text{при} \quad (m, n) \in \{\hat{I}_a \cap I_d\}.$$

Поэтому d_{mn} можно ввести под знак двойной суммы; тогда получим

$$e_{mn} = \sum_{\substack{i=1 \\ i \cdot j \neq 1}}^{M_b} \sum_{j=1}^{N_b} b_{ij} d_{m-i+1,\, n-j+1} \quad \text{при} \quad (m, n) \in \{\hat{I}_a \cap I_d\}.$$

Поскольку мы определили ошибку, можно выбрать b_{ij} таким образом, чтобы минимизировалась среднеквадратичная ошибка:

$$\overline{e^2} = \sum_{(m, n)} \sum \left[\sum_{(i, j)} b_{ij} d_{m-i+1,\, n-j+1} \right]^2 \quad \text{при} \quad (i, j) \in I_b,$$
$$(m, n) \in \{\hat{I}_a \cap I_d\}. \quad (4.34)$$

Если не считать b_{11}, равного единице, коэффициенты b_{ij} еще предстоит определить. Продифференцировав (4.34) по b_{ij} и приравняв полученное выражение нулю, можно найти значения b_{ij}, минимизирующие среднеквадратичную ошибку. В результате получим $(M_b \cdot N_b - 1)$ уравнений вида

$$\sum_{\substack{i=1 \\ i \cdot j \neq 1}}^{M_b} \sum_{j=1}^{N_b} b_{ij} \Phi_{klij} = \Phi_{kl} \quad \text{при} \quad k = 1, 2, 3, \ldots, M_b, \quad l = 1, 2, 3, \ldots, N_b, \quad (4.35)$$
$$\text{причем } k \cdot l \neq 1;$$

здесь

$$\Phi_{klij} = \sum_{(m,\,n)} d_{m-l+1,\,n-j+1} d_{m-k+1,\,n-i+1}, \tag{4.36}$$

$$\Phi_{kl} = -\sum_{(m,\,n)} d_{mn} d_{m-k+1,\,n-l+1}, \quad (m,\,n) \in \{\hat{I}_a \cap I_d\}. \tag{4.37}$$

Таким образом, (4.35) представляет собой систему $(M_b \cdot N_b - 1)$ совместных линейных уравнений с $(M_b \cdot N_b - 1)$ неизвестными. Решения уравнений (4.35) дают коэффициенты b_{ij}, минимизирующие среднеквадратичную ошибку.

После определения коэффициентов знаменателя $B(z_1, z_2)$ передаточной функции фильтра остается вычислить коэффициенты числителя $A(z_1, z_2)$. Один из возможных подходов состоит в нахождении таких коэффициентов $A(z_1, z_2)$, которые минимизировали бы среднеквадратичное значение разности между коэффициентами передаточной функции $F(z_1, z_2) = A(z_1, z_2)/B(z_1, z_2)$ и коэффициентами требуемой передаточной функции $D(z_1, z_2)$. Это задача дискретной винеровской фильтрации в двух измерениях [20]. Задача заключается в нахождении оптимальной фильтрующей функции $A(z_1, z_2)$ при заданном входном сигнале $1/B(z_1, z_2)$ и желаемом выходном сигнале $D(z_1, z_2)$.

Более простой, но менее точный метод нахождения массива A состоит в свертке массивов B и D. Поскольку $A(z_1, z_2)/B(z_1, z_2) \approx D(z_1, z_2)$, коэффициенты a_{mn} можно вычислить из соотношения $A(z_1, z_2) = B(z_1, z_2) \cdot D(z_1, z_2)$ при $(m, n) \in I_a$. Если $B(z_1, z_2)$ выбрано удачно, коэффициенты $B(z_1, z_2)$ будут иметь относительно небольшую величину при $(m, n) \in I_a$.

Пример

Для иллюстрации описанного метода рассмотрим импульсную характеристику двумерного рекурсивного фильтра с передаточной функцией $D(z_1, z_2) = A'(z_1, z_2)/B'(z_1, z_2)$, где A' и B' — массивы размерности 3×3. Значения коэффициентов A' и B' представлены в табл. 4.1 (столбцы A' и B'). В качестве желательной импульсной характеристики использовались 20×20 ее первых значений. Для M_a, M_b, N_a и N_b было принято значение 3. После вычисления коэффициентов согласно (4.36) и (4.37) была решена система уравнений (4.35), в результате чего были получены коэффициенты знаменателя. Коэффициенты числителя были определены путем двумерной свертки $B(z_1, z_2)$ и $D(z_1, z_2)$. Коэффициенты рассматриваемого рекурсивного фильтра, вычисленные с помощью машинной программы, представлены в табл. 4.1 (столбцы A и B).

Полученные результаты свидетельствуют о весьма хорошем соответствии с коэффициентами исходного фильтра. Однако вы-

Таблица 4.1

Проверка синтезированной импульсной характеристики

Коэффициент	A'	A	B'	B
Константа	1,0	1,0	1,0	1,0
z_1	3,0	3,0000005	—1,2	—1,1999995
z_1^2	2,0	2,0000016	0,5	0,4999996
z_2	2,0	2,0000002	—1,5	—1,4999998
z_1z_2	4,0	4,0000014	1,8	1,7999989
$z_1^2z_2$	—1,0	—0,9999976	—0,75	—0,7499992
z_2^2	—1,0	—0,9999995	0,6	0,5999998
$z_1z_2^2$	2,0	2,0000002	—0,72	—0,7199994
$z_1^2z_2^2$	1,0	0,9999998	0,29	0,2899996

Примечание. В столбцах A' и B' даны фактические коэффициенты, а в столбцах A и B — аппроксимированные коэффициенты.

бор правильного порядка массивов числителя и знаменателя представляет собой сложную задачу, даже если речь идет об одномерном случае. В данном примере сразу были выбраны «правильные» порядки, поскольку было заранее известно, что искомая импульсная характеристика связана с фильтром, имеющим передаточную функцию с числителем и знаменателем размерности 3×3. В более общем случае, когда такая информация отсутствует, наиболее подходящим является метод проб и ошибок.

4.3.3. Обобщенный подход к аппроксимации и устойчивости

Аппроксимация

Аппроксимационная задача общего вида, которую приходится решать при проектировании двумерных рекурсивных фильтров, является задачей нелинейной минимизации. Задача сводится к выбору таких коэффициентов массивов A и B, чтобы частотная характеристика фильтра была близкой аппроксимацией требуемой частотной характеристики. Часто используют критерий ошибки аппроксимации в виде среднеквадратичной нормы. Имеется множество алгоритмов, позволяющих осуществить минимизацию подобного рода. Например, можно указать работы [21, 22, 27].

Методы обеспечения устойчивости

Если на коэффициенты фильтра не накладывать никаких ограничений, то фильтр с фактически определенными коэффициентами может оказаться неустойчивым. Существуют два возможных метода обеспечения устойчивости, которые описываются ниже. Оба метода предполагают проведение определенных операций над полиномом знаменателя, направленных на изменение его фазочастотной характеристики (ФЧХ). Они позволяют создать фильтр, АЧХ которого подобна АЧХ исходного неустойчивого фильтра.

Двойная двумерная инверсия по критерию наименьших квадратов. Первый метод основан на использовании свойств двумерной инверсии по критерию наименьших квадратов. Предположим, что задан некоторый массив C. Требуется найти такой массив P, что C после свертки с P аппроксимирует массив единичного импульса U. Таким образом, имеем

$$C * P \approx U. \tag{4.38}$$

Символ * обозначает свертку.

Определение двумерной инверсии. В общем случае невозможно обеспечить точное равенство $C * P$ и U. Фактически оказывается, что результат $C * P$ равен другому массиву G. Если P выбрать таким образом, чтобы минимизировалась сумма квадратов элементов $U - G$, то такое P будет называться *инверсией по критерию наименьших квадратов* C [24]. Если C, P, G и U — двумерные массивы, то в этом случае P называется двумерной инверсией по критерию наименьших квадратов (ДИНК) C. В двумерном случае P находится по U и C с помощью двумерного метода Винера, описанного Уиггинсом [20].

Предположение о минимальной фазе. Двумерные инверсии по критерию наименьших квадратов как группа обладают рядом интересных свойств. Одно такое свойство, которым мы будем пользоваться, рассматривается в следующем предположении.

Предположение А. Если задана произвольная действительная конечномерная матрица C, то любая ее двумерная инверсия по критерию наименьших квадратов обладает минимально-фазовыми свойствами. Смысл этого предположения сводится к тому, что при задании матрицы C, по которой можно определить ДИНК-матрицу P, последняя обязательно должна быть минимально-фазовой.

Это очень важное предположение, поскольку оно означает, что фильтр с передаточной функцией $F(z_1, z_2) = 1/P(z_1, z_2)$ окажется устойчивым, если P есть ДИНК.

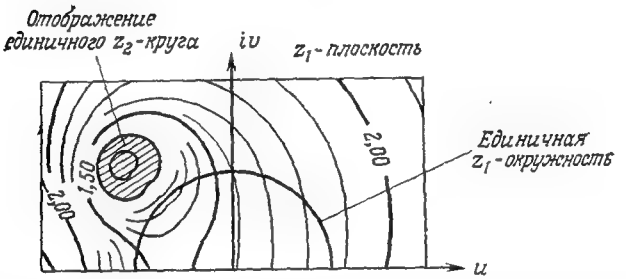
а

$$\begin{bmatrix} 1,0 & -1,2 & 0,4 \\ -1,2 & 1,7 & -0,65 \\ 0,4 & -0,7 & 0,25 \end{bmatrix}$$

б

$$\begin{bmatrix} 1,0 & 0,87742 & 0,39488 \\ 0,90441 & 0,65133 & 0,22754 \\ 0,39818 & 0,21555 & 0,04207 \end{bmatrix}$$

в



Фиг. 4.21. Матрица входного массива A (*а*). ДИНК размерности 3×3 массива A, где ДИНК — двумерная инверсия по критерию наименьших квадратов (*б*). Корневая диаграмма ДИНК размерности 3×3 массива A (*в*).

К сожалению, мы не располагаем доказательством этого предположения, которое является обобщением одномерной теоремы [24]. Однако нами было проверено множество численных примеров, и не оказалось ни одного случая, который бы свидетельствовал об обратном [1]).

Пример. В качестве примера использования ДИНК рассмотрим матрицу A, приведенную на фиг. 4.21, *а*. ДИНК размерности 3×3 матрицы A представлена на фиг. 4.21, *б*. Корневая диаграмма этой ДИНК показана на фиг. 4.21, *в*. Поскольку отображение единичного z_2-круга (заштрихованная область) не пересекается единичной z_1-окружностью, ДИНК должна быть минимально-фазовой.

Предположим, что был спроектирован двумерный рекурсивный фильтр с передаточной функцией $F(z_1, z_2) = 1/B(z_1, z_2)$, который, однако, оказался неустойчивым; другими словами,

[1]) Т. Хуанг, частное сообщение (март 1971 г.).

знаменатель передаточной функции оказался неминимально-фазовым. Каким образом следует изменить коэффициенты, чтобы получить устойчивый фильтр?

Обозначим матрицу знаменателя передаточной функции неустойчивого фильтра через B. Вычислим B' — двумерную инверсию по критерию наименьших квадратов B. Из предположения А следует, что B' (ДИНК B) является минимально-фазовой. Предположим, что мы вычислили ДИНК B' и обозначили ее $\hat{B}$. Таким образом, $\hat{B}$ есть инверсия инверсии, т. е. «двойная инверсия» B. Интуиция подсказывает, что $\hat{B}$ и B обладают некоторыми общими свойствами. Более того, $\hat{B}$, как таковая, есть ДИНК. Следовательно, $\hat{B}$ является минимально-фазовой, а B нет. Поэтому фильтр с передаточной функцией $\hat{F}(z_1, z_2) = 1/\hat{B}(z_1, z_2)$ устойчив.

Рассмотрим, например, фильтр с передаточной функцией $F(z_1, z_2) = 1/B(z_1, z_2)$, где

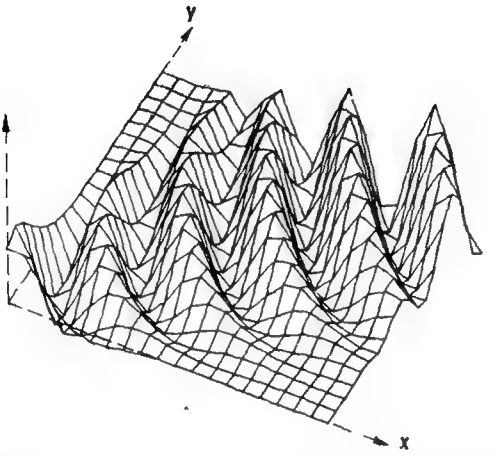
$$B = \begin{bmatrix} 1{,}0 & -1{,}20002759 & 0{,}40002239 \\ -1{,}00003018 & 1{,}70007079 & -0{,}65005088 \\ 0{,}40002035 & -0{,}7005488 & 0{,}25004387 \end{bmatrix}. \quad (4.39)$$

Этот фильтр неустойчив, что иллюстрирует фиг. 4.22, где показан перспективный вид импульсной характеристики, соответствующей $F(z_1, z_2)$. Следует отметить возрастание уровня импульсной характеристики по мере увеличения значений пространственных координат x и y. Убедиться в неустойчивости фильтра можно путем построения корневой диаграммы B. Последняя представлена на фиг. 4.23. Поскольку заштрихованная область (отображение единичного z_2-круга) частично пересекается единичной z_1-окружностью, B не является минимально-фазовым.

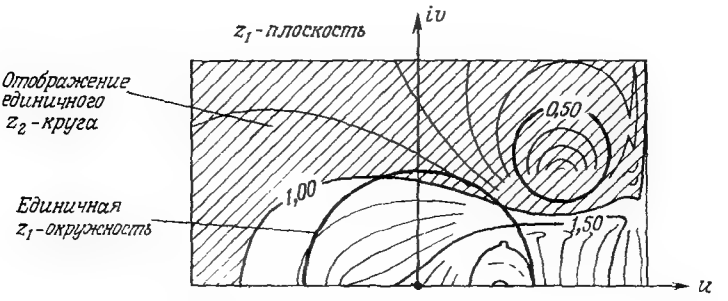
Чтобы получить устойчивый фильтр, прежде всего следует вычислить B', т. е. ДИНК B. В рассматриваемом примере B' выражается матрицей размера 3×3. Вычислим ДИНК-матрицу размера 3×3, которую обозначим через $\hat{B}$:

$$\hat{B} = \begin{bmatrix} 1{,}11567 & -1{,}23469 & 0{,}38431 \\ -1{,}0932 & 1{,}68765 & -0{,}6821 \\ 0{,}37577 & -0{,}65178 & 0{,}29767 \end{bmatrix}. \quad (4.40)$$

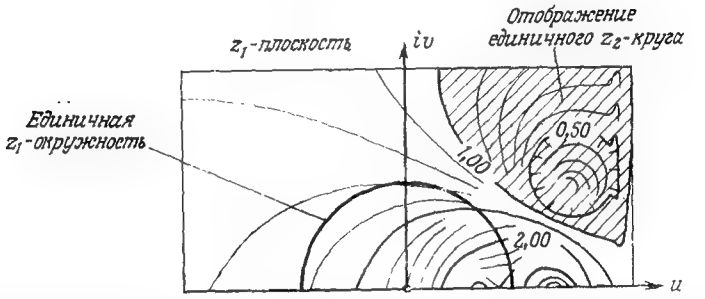
Корневая диаграмма $\hat{B}$ представлена на фиг. 4.24, которая показывает, что $\hat{B}$ является минимально-фазовым. Следовательно, фильтр с передаточной функцией $\hat{F}(z_1, z_2) = 1/\hat{B}(z_1, z_2)$ должен быть устойчивым. Это заключение подтверждает изучение фиг. 4.25, на которой показан перспективный вид импульсной характеристики, соответствующей $\hat{F}(z_1, z_2)$. Очевидно, что уро-



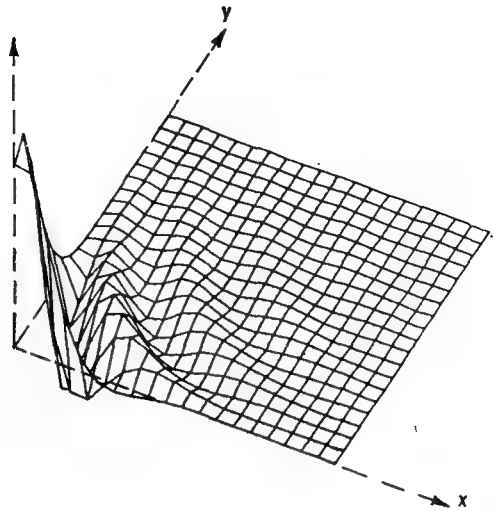
Фиг. 4.22. Двумерная импульсная характеристика неустойчивого фильтра.



Фиг. 4.23. Корневая диаграмма неустойчивого фильтра.



Фиг. 4.24. Корневая диаграмма двойной ДИНК знаменателя неустойчивой передаточной функции.



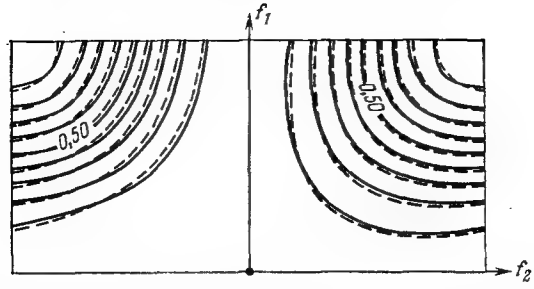
Фиг. 4.25. Двумерная импульсная характеристика фильтра после обеспечения его устойчивости.

вень этой импульсной характеристики снижается по мере увеличения координат x и y.

Замечания относительно амплитудного и фазового спектров. В каком смысле B и $\hat{B}$ являются подобными? Как мы уже убедились, $\hat{B}$ является минимально-фазовым, а B таким свойством не обладает. С другой стороны, $\hat{B}$ есть двойная ДИНК B. Поэтому мы предполагаем, что фактически $\hat{B}$ должно быть аппроксимирующим минимально-фазовым вариантом B. Другими словами, можно ожидать, что амплитудные спектры B и $\hat{B}$ являются приблизительно одинаковыми. Двумерные амплитудные спектры B и $\hat{B}$ представлены на фиг. 4.26 в виде контурных диаграмм. Описанный метод позволил создать устойчивый фильтр путем модификации знаменателя передаточной функции неустойчивого фильтра. Оба массива знаменателя обладают схожими амплитудными спектрами, однако фазовый спектр B был видоизменен для обеспечения устойчивости.

Двумерное дискретное преобразование Гильберта. Второй метод обеспечения устойчивости двумерных рекурсивных фильтров называется «методом преобразования Гильберта» [25]. Этот метод, как и предыдущий, основан на обобщении одномерной методики.

Одномерный случай. Как известно [26], в случае минимально-фазовой последовательности $a(i)$ фазовый спектр $\theta(\exp j\omega)$



Фиг. 4.26. Сравнение двух амплитудных спектров неминимально-фазового знаменателя B (штриховые линии) и минимально-фазового знаменателя $\hat{B}$ (сплошные линии).

и логарифм амплитудного спектра $\log|A(\exp j\omega)|$ связаны между собой преобразованием Гильберта:

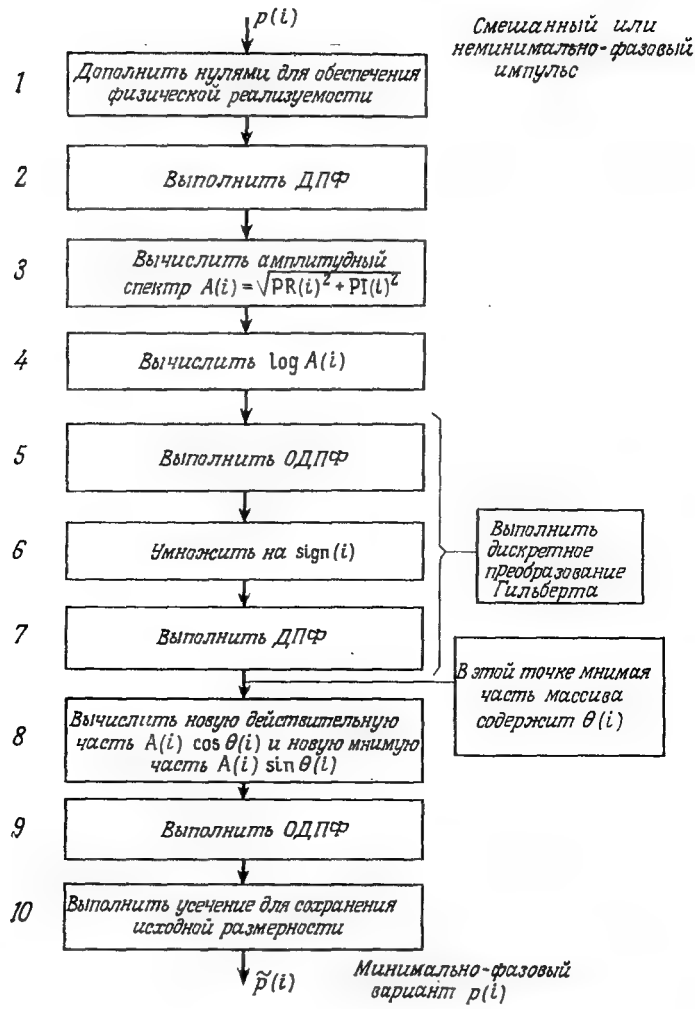
$$\theta(\exp j\omega) = -\frac{1}{2\pi}\int_0^{2\pi} \log|A(\exp j\Omega)| \operatorname{ctg}\frac{\omega-\Omega}{2}\, d\Omega. \qquad (4.41)$$

Соотношение (4.41) вытекает из следующего факта: последовательность $a(i)$ $(i = 0, 1, \ldots, M-1)$ является минимально-фазовой, если и только если обратное z-преобразование $\log[A(z)]$ является физически реализуемым. Поскольку преобразование Гильберта связывает действительную и мнимую части физически реализуемых функций, применение такого преобразования к $\log|A(\exp j\omega)|$ [соотношение (4.41)] дает мнимую часть $\log[A(z)]$, которая является фазовым спектром $A(z)$.

Используя правило трапеций, можно найти дискретную аппроксимацию (4.41) на основе дискретного преобразования Фурье [25]. Процедура нахождения аппроксимированного минимально-фазового спектра по известному амплитудному спектру смешанной или минимально-фазовой последовательности $a(i)$ определяется соотношением

$$\theta(i) = -j\,\text{ДПФ}\{\text{sign}(i)\cdot\text{ОДПФ}[\log A(i)]\}. \qquad (4.42)$$

Здесь $A(i)$ — i-я составляющая амплитудного спектра, $\theta(i)$ — i-я составляющая фазового спектра. ДПФ и ОДПФ выражают выполнение дискретного преобразования Фурье и обратного дискретного преобразования Фурье. Функция $\text{sign}(i)$ есть конечномерная дискретная функция знака, длина которой равна длине интервала выполнения дискретного преобразования Фурье. На фиг. 4.27 представлена блок-схема алгоритма для получения минимально-фазового варианта смешанного или неминимально-фазового импульса.



Фиг. 4.27. Блок-схема алгоритма для получения минимально-фазового варианта одномерного импульса.

Двумерный случай. Если воспользоваться двумерным вариантом дискретного преобразования Фурье и найти особую двумерную функцию знака, то рассматриваемый метод можно обобщить на двумерный случай [25]. Таким образом, при заданном двумерном амплитудном спектре $A(i_1, i_2)$ физически реализуемой последовательности минимально-фазовый спектр

$\theta(i_1, i_2)$ вычисляется с помощью выражения

$$\theta(i_1, i_2) = -j\,\text{ДПФ}\{\text{sign}(i_1, i_2) \cdot \text{ОДПФ}[\log A(i_1, i_2)]\}. \quad (4.43)$$

Блок-схема алгоритма для получения минимально-фазового варианта смешанного или неминимально-фазового массива, соответствующего (4.43), не отличается от алгоритма, приведенного на фиг. 4.27, за исключением того, что все операции необходимо выполнять в двух измерениях.

Сравнение с методом двойной двумерной инверсии по критерию наименьших квадратов

Пример 1. В первом примере метод обеспечения устойчивости с применением преобразования Гильберта сравнивается с методом двойной двумерной инверсии по критерию наименьших квадратов. Рассматривается двумерный рекурсивный фильтр, в знаменателе передаточной функции которого используется полином $B(z_1, z_2)$ с коэффициентами

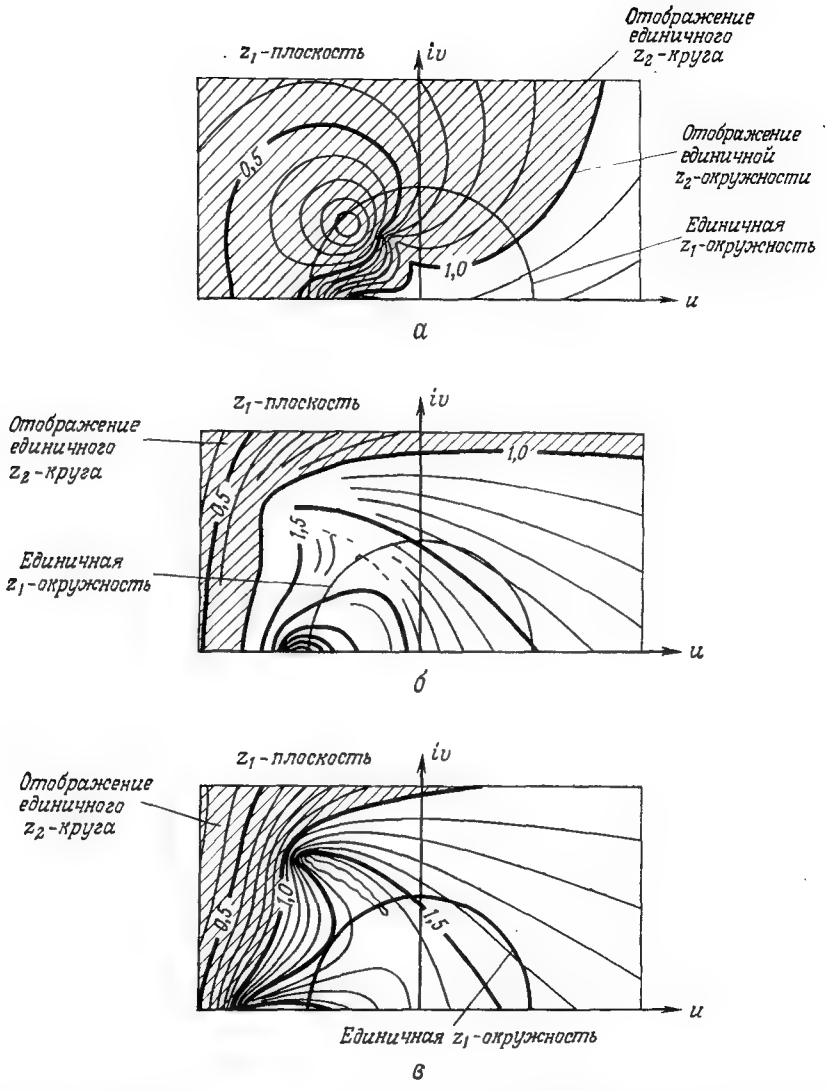
$$B = \begin{bmatrix} 1{,}0 & -0{,}75 & 0{,}9 \\ 1{,}5 & -1{,}2 & 1{,}3 \\ 1{,}2 & 0{,}9 & 0{,}5 \end{bmatrix}. \quad (4.44)$$

Корневая диаграмма для этого массива коэффициентов приведена на фиг. 4.28, *а*. Видно, что отображение единичного z_2-круга в z_1-плоскости пересекается единичной z_1-окружностью. Это означает, что фильтр с полиномом $B(z_1, z_2)$ в знаменателе его передаточной функции неустойчив. На фиг. 4.28, *б* показана корневая диаграмма массива

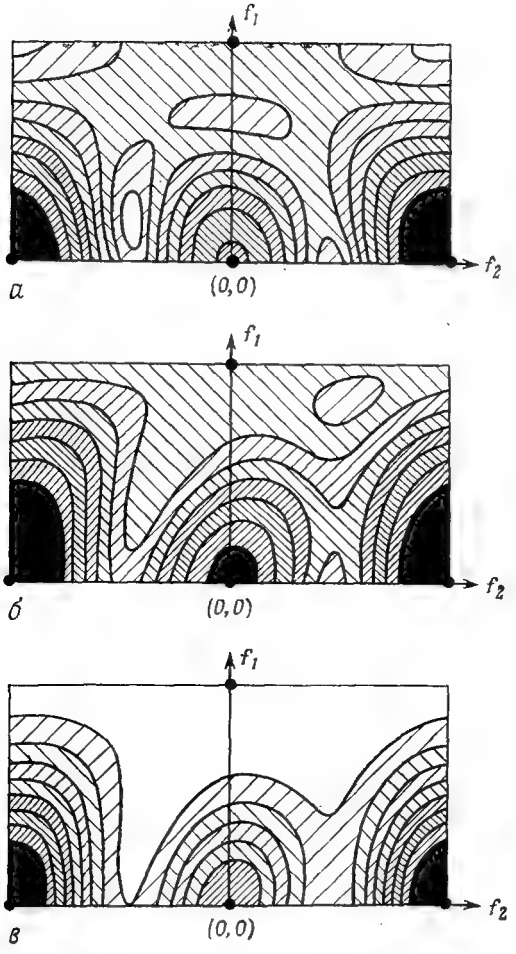
$$G = \begin{bmatrix} 6{,}47 & -0{,}139 & -1{,}81 \\ 2{,}98 & -0{,}767 & 2{,}04 \\ 0{,}380 & 0{,}981 & -0{,}466 \end{bmatrix}, \quad (4.45)$$

который получен из массива B в результате применения метода двойной инверсии по методу наименьших квадратов. Теперь отображение единичного z_2-круга в z_1-плоскости не пересекается единичной z_1-окружностью. Это означает, что рекурсивный фильтр с полиномом $G(z_1, z_2)$ в знаменателе передаточной функции устойчив. На фиг. 4.28, *в* показана корневая диаграмма двумерного полинома с коэффициентами

$$H = \begin{bmatrix} 2{,}46 & -0{,}619 & 0{,}881 \\ 1{,}29 & -0{,}427 & 0{,}770 \\ 2{,}63 & 0{,}325 & 0{,}212 \end{bmatrix}. \quad (4.46)$$



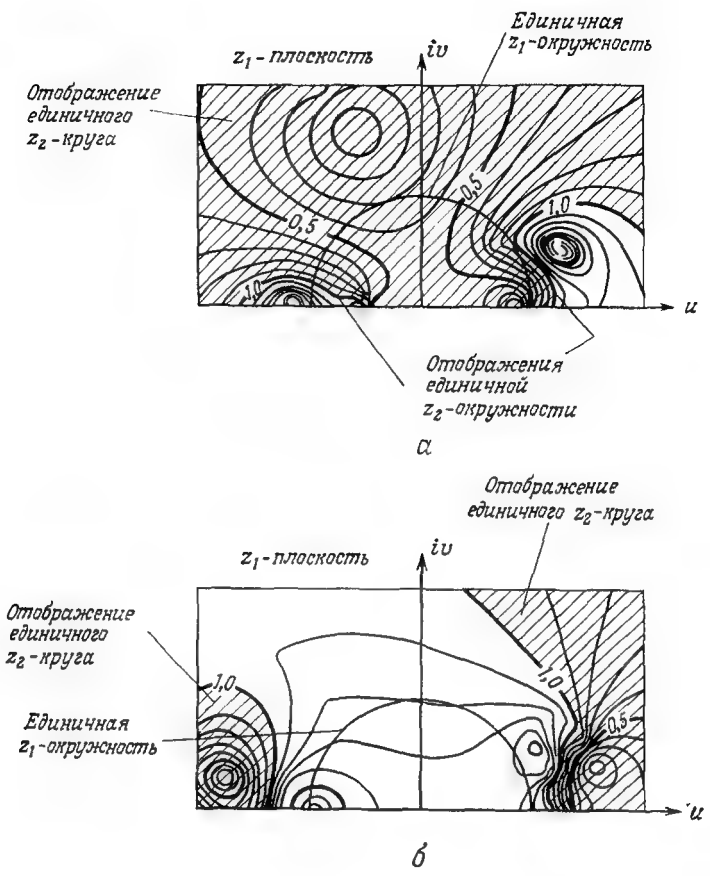
Фиг. 4.28. Корневая диаграмма для неминимально-фазового двумерного массива (*а*), корневая диаграмма для варианта неминимально-фазового двумерного массива, полученного после обеспечения устойчивости методом двойной ДИНК (*б*), корневая диаграмма для варианта неминимально-фазового двумерного массива, полученного после обеспечения устойчивости методом преобразования Гильберта.



Фиг. 4.29. Сравнение амплитудных спектров, полученных после применения различных методов обеспечения устойчивости: *а* — контурная диаграмма исходного амплитудного спектра; *б* — контурная диаграмма после применения метода преобразования Гильберта; *в* — контурная диаграмма после применения метода двойной ДИНК.

Этот массив получен из массива коэффициентов B методом обеспечения устойчивости с применением преобразования Гильберта.

Корневая диаграмма $H(z_1, z_2)$ показывает, что отображение единичного z_2-круга не пересекается единичной z_1-окружностью в z_1-плоскости. Следовательно, рекурсивный фильтр с



Фиг. 4.30. Корневая диаграмма неминимально-фазового двумерного массива (*а*), та же корневая диаграмма после обеспечения устойчивости методом преобразования Гильберта (*б*).

полиномом H в знаменателе передаточной функции является минимально-фазовым. Различная конфигурация контуров равного уровня на фиг. 4.28, *б* и *в* свидетельствует о том, что двумерные спектры H и G существенно отличаются друг от друга. Контурные диаграммы амплитудных спектров B, H и G представлены на фиг. 4.29, *а*—*в*. Квадрат ошибки для получаемого амплитудного спектра определяется выражением

$$e(B, G) = \sum_{i_1=0}^{N_1-1} \sum_{i_2=0}^{N_2-1} [AB(i_1, i_2) - AG(i_1, i_2)]^2, \qquad (4.47)$$

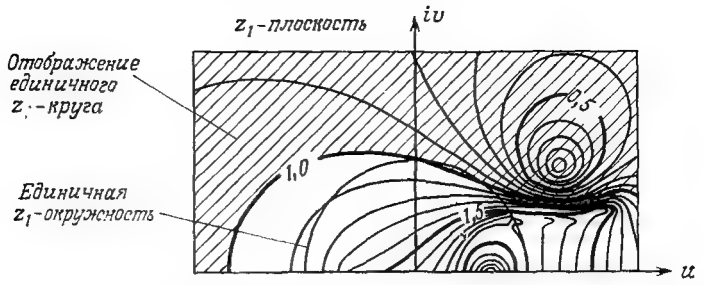
где AB и AG — амплитудные спектры B и G соответственно. Ошибка $e(B, H)$ определяется аналогично. В случае амплитудных спектров на фиг. 4.29 ошибка $e(B, H)$ в 18,5 раза меньше ошибки $e(B, G)$.

Из блок-схемы на фиг. 4.27 следует, что усечение массива H производится после того, как H определяется непосредственно из обратного дискретного преобразования Фурье AH и из фазового спектра, полученного с помощью дискретного преобразования Гильберта. Усечение полученного результата производится таким образом, чтобы восстанавливалась размерность исходного массива B. Если применение метода обеспечения устойчивости с дискретным преобразованием Гильберта дает абсолютно точный минимально-фазовый массив, то все элементы обратного преобразования AH, размерность которых превышает размерность исходного массива B, имеют нулевое значение. Практически же величина элементов H вне области, границы которой определяются размерностью исходного массива B, оказывается малой по сравнению с величиной элементов в пределах этой области, но не равной нулю. Эти элементы имеют ненулевые значения потому, что метод дискретного преобразования Гильберта, реализуемый с применением дискретного преобразования Фурье, представляет собой аппроксимацию интегрального преобразования [25]. Метод преобразования Гильберта может привести к получению массивов, не являющихся минимально-фазовыми. Однако значительный вычислительный опыт показывает, что такая ситуация не должна возникать часто; в большинстве случаев рассматриваемый метод дает минимально-фазовый вариант исходного массива с амплитудным спектром, близко аппроксимирующим амплитудный спектр исходного массива.

Пример 2. Полином $B(z_1, z_2)$ с коэффициентами, задаваемыми матрицей размера 5×5

$$B = \begin{bmatrix} 1{,}00 & 1{,}50 & -1{,}90 & -0{,}80 & 1{,}10 \\ 1{,}40 & 2{,}10 & -2{,}60 & -1{,}10 & 1{,}50 \\ -1{,}80 & -2{,}40 & 3{,}30 & 1{,}30 & -1{,}60 \\ -0{,}70 & -0{,}90 & 1{,}10 & 0{,}50 & -0{,}80 \\ -0{,}90 & 1{,}30 & -1{,}60 & -0{,}60 & 1{,}00 \end{bmatrix}, \qquad (4.48)$$

имеет корневую диаграмму, показанную на фиг. 4.30, *а*. Как и в предыдущем примере, здесь отображение единичного z_2-круга пересекается единичной z_1-окружностью в z_1-плоскости. Таким образом, рекурсивный фильтр с $B(z_1, z_2)$ в знаменателе передаточной функции неустойчив. Применение метода обеспечения



Фиг. 4.31. Корневая диаграмма, полученная в результате применения метода обеспечения устойчивости с преобразованием Гильберта в контрпримере.

устойчивости с преобразованием Гильберта приводит к получению минимально-фазового массива H вида

$$H = \begin{bmatrix} 4{,}76 & -0{,}891 & -3{,}43 & 0{,}851 & 0{,}847 \\ -0{,}392 & 0{,}206 & 0{,}305 & 0{,}0787 & -0{,}0187 \\ -3{,}04 & 0{,}721 & 2{,}94 & -0{,}753 & -0{,}853 \\ 0{,}204 & -0{,}469 & -0{,}503 & 0{,}238 & 0{,}165 \\ 0{,}586 & -0{,}416 & -0{,}704 & 0{,}312 & 0{,}211 \end{bmatrix}, \quad (4.49)$$

корневая диаграмма которого показана на фиг. 4.30, *б*. Отображение единичного z_2-круга не пересекается единичной z_1-окружностью, т. е. фильтр с $H(z_1, z_2)$ в знаменателе передаточной функции устойчив. Следует обратить внимание на то, что точка возмущения с координатами приблизительно $z_1 = (-0{,}5, 0)$ на фиг. 4.30, *а* отразилась относительно единичной z_1-окружности и отобразилась в точку с координатами приблизительно $z_1 = (-2, 0)$ на фиг. 4.30, *б*.

Контрпример и замечания. Метод обеспечения устойчивости с преобразованием Гильберта, позволяющий находить очень хорошие аппроксимации в спектральной области, не лишен, однако, недостатков. Приложение этого метода к массиву (4.39) приводит к получению массива, корневая диаграмма которого показана на фиг. 4.31. Видно, что единичная z_1-окружность пересекает отображение единичного z_2-круга несмотря на применение метода обеспечения устойчивости. В данном случае метод с преобразованием Гильберта оказался непригодным для получения минимально-фазового массива. Возможны два объяснения этого факта: 1) метод теоретически не гарантирует нахождения минимально-фазового массива и 2) имеются вычислительные трудности, проявившиеся в данном примере. Корневая диаграмма на фиг. 4.23 показывает, что контурная линия с уровнем

1,00 проходит на очень близком расстоянии от единичной z_1-окружности в зоне, где эта единичная окружность пересекает отображение единичного z_2-круга. Фиг. 4.31 показывает, что метод с преобразованием Гильберта уменьшил, но не ликвидировал зону пересечения. Описанное явление можно также объяснить вычислительными трудностями, которые возникают при работе с одномерными массивами, нули z-преобразования которых близки по величине к единице.

4.4. Реализация

После завершения проектирования и обеспечения устойчивости двумерного рекурсивного фильтра в него необходимо ввести данные. Как следует из (4.3), можно разработать рекурсивный алгоритм, описывающий рекурсивное соотношение в случае фильтра с рекурсией в направлении положительных m и n.

4.4.1. Фазочастотная характеристика двумерного рекурсивного фильтра

Алгоритм рекурсивной фильтрации (4.3) позволяет создать фильтр с «одноквадрантной» импульсной характеристикой. Если входным сигналом является единичный импульс, заданный на некоторой решетке точек, то фильтр реагирует только в области положительных значений m и n. Поскольку импульсная характеристика рекурсивного фильтра является одноквадрантной, то он не может обладать нулевой ФЧХ. Однако путем соответствующего сочетания двух или больше рекурсивных фильтров можно обеспечить фильтрацию, которая характеризуется нулевой ФЧХ или другими полезными симметричными разновидностями ФЧХ.

Получение нулевой ФЧХ методом последовательного соединения фильтров

Двумерный фильтр с нулевой ФЧХ обладает тем свойством, что его импульсная характеристика симметрична относительно начала координат (x, y) по любому направлению, проходящему через начало координат. Имеется возможность осуществить фильтрацию с нулевой ФЧХ, используя рекурсивный фильтр. Для этого требуется сначала профильтровать входные данные с помощью рекурсивного фильтра, начиная с одного из углов массива этих данных. Затем необходимо профильтровать полученный массив с помощью того же рекурсивного фильтра, но начиная с противолежащего угла массива и изменив направление фильтрации на противоположное [10]. Эти операции

эквивалентны фильтрации с помощью фильтра, имеющего передаточную функцию

$$G(z_1, z_2) = F(z_1, z_2) \cdot F(1/z_1, 1/z_2), \tag{4.50}$$

где $F(z_1, z_2)$ — передаточная функция одноквадрантного рекурсивного фильтра [1]).

При $z_1 = \exp(j\omega_1)$ и $z_2 = \exp(j\omega_2)$ выражение (4.50) позволяет найти двумерную частотную характеристику, соответствующую передаточной функции $G(z_1, z_2)$. Поскольку $F(e^{j\omega_1}, e^{j\omega_2})$ является комплексно-сопряженной функцией относительно $F(e^{-j\omega_1}, e^{-j\omega_2})$, $G(e^{j\omega_1}, e^{j\omega_2})$ всегда будет действительной функцией при любых значениях (ω_1, ω_2). Поэтому фильтр с передаточной функцией $G(z_1, z_2)$ имеет нулевую ФЧХ. Соответствующая этой передаточной функции АЧХ равна квадрату АЧХ, соответствующей $F(z_1, z_2)$. Таким образом, $|G(\omega_1, \omega_2)| = |F(\omega_1, \omega_2)|^2$.

Получение нулевой ФЧХ методом суммирования

Нулевую ФЧХ можно получить в том случае, если профильтровать входной массив в прямом и обратном направлениях и сложить полученные результаты. Это эквивалентно применению фильтра с передаточной функцией вида

$$H(z_1, z_2) = F(z_1, z_2) + F(1/z_1, 1/z_2). \tag{4.51}$$

Как и в случае $G(z_1, z_2)$, $H(z_1, z_2)$ имеет нулевую ФЧХ. Однако этим передаточным функциям соответствуют различные АЧХ. Для $H(z_1, z_2)$ АЧХ имеет вид

$$|H(\omega_1, \omega_2)| = 2\,\mathrm{Re}\{F(\omega_1, \omega_2)\}. \tag{4.52}$$

Фильтрацию можно сделать симметричной относительно оси x (или оси y), если профильтровать данные в прямом и обратном x-направлениях (или y-направлениях). Например, если z_1 соответствует оси x, а z_2 — оси y, то $G(z_1, z_2) = F(z_1, z_2) \cdot F(1/z_1, z_2)$ и $H(z_1, z_2) = F(z_1, z_2) + F(1/z_1, z_2)$ будут симметричными относительно оси x.

Наконец, фильтрации вида

$$G(z_1, z_2) = F(z_1, z_2) \cdot F(1/z_1, z_2) \cdot F(z_1, 1/z_2) \cdot F(1/z_1, 1/z_2) \tag{4.53}$$

и

$$H(z_1, z_2) = F(z_1, z_2) + F(1/z_1, z_2) + F(z_1, 1/z_2) + F(1/z_1, 1/z_2) \tag{4.54}$$

симметричны относительно обеих осей x и y и обеспечивают нулевую ФЧХ.

[1]) Применение этого метода требует достаточно большой длины выходного массива ЦВМ, чтобы выходные данные могли достигать относительно малых численных значений.

4.4.2. Реализация путем применения переменных состояния [1])

Если воспользоваться массивом переменных состояния, то можно получить реализацию, отличающуюся от реализации прямым методом, согласно (4.3). Этот альтернативный метод реализации рекурсивного алгоритма может обеспечить уменьшение требуемого объема памяти ЦВМ. Массив переменных состояния обозначим $W(z_1, z_2)$. Это отражено в соотношении

$$\frac{R(z_1, z_2)}{D(z_1, z_2)} = \frac{A(z_1, z_2)\, W(z_1, z_2)}{B(z_1, z_2)\, W(z_1, z_2)}. \tag{4.55}$$

Чтобы проследить, каким образом введение этого массива сказывается на реализации, сделаем отождествления

$$D(z_1, z_2) = B(z_1, z_2)\, W(z_1, z_2), \tag{4.56}$$

$$R(z_1, z_2) = A(z_1, z_2)\, W(z_1, z_2). \tag{4.57}$$

Уравнения (4.56) и (4.57) предполагают выполнение двух операций. Во-первых, массив B воздействует рекурсивным образом на массив D, в результате чего получается массив переменных состояния W; во-вторых, массив A воздействует на массив W, в результате чего вычисляется выходной массив R.

Уравнение (4.56) дает (принимается, что операции нормализации, согласно (4.3), уже выполнены)

$$\left(1 + \sum_{\substack{k=1 \\ k \cdot l \neq 1}}^{M_b} \sum_{l=1}^{N_b} b_{kl} z_1^{k-1} z_2^{l-1}\right) W(z_1, z_2) = D(z_1, z_2), \tag{4.58}$$

$$W(z_1, z_2) = D(z_1, z_2) - W(z_1, z_2) \sum_{\substack{k=1 \\ k \cdot l \neq 1}}^{M_b} \sum_{l=1}^{N_b} b_{kl} z_1^{k-1} z_2^{l-1}. \tag{4.59}$$

Рекурсивные уравнения

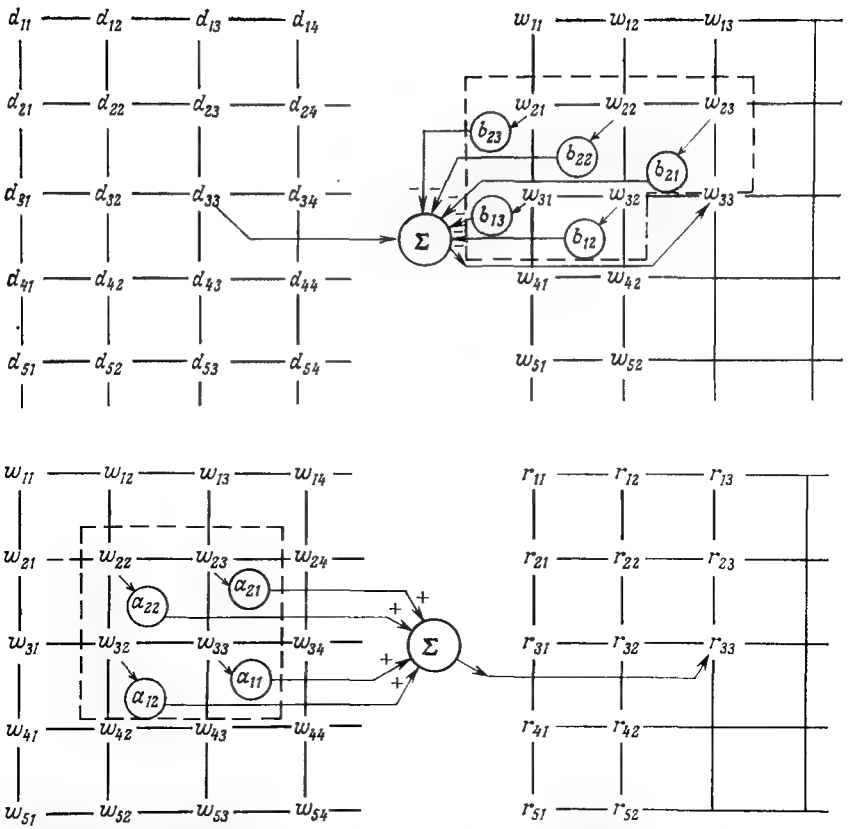
Из уравнения (4.59) с очевидностью следует рекурсивный характер алгоритма. Преобразование, обратное (4.59), имеет вид

$$w_{mn} = d_{mn} - \sum_{\substack{k=1 \\ k \cdot l \neq 1}}^{M_b} \sum_{l=1}^{N_b} b_{kl} w_{m-k+1,\, n-l+1}. \tag{4.60}$$

Уравнение (4.60) описывает эволюцию массива переменных состояния. Выходной массив R формируется из массива переменных состояния:

$$r_{mn} = \sum_{i=1}^{M_a} \sum_{j=1}^{N_a} a_{ij} w_{m-i+1,\, n-j+1}. \tag{4.61}$$

[1]) Этот материал принадлежит одному из авторов (Р. Риду).

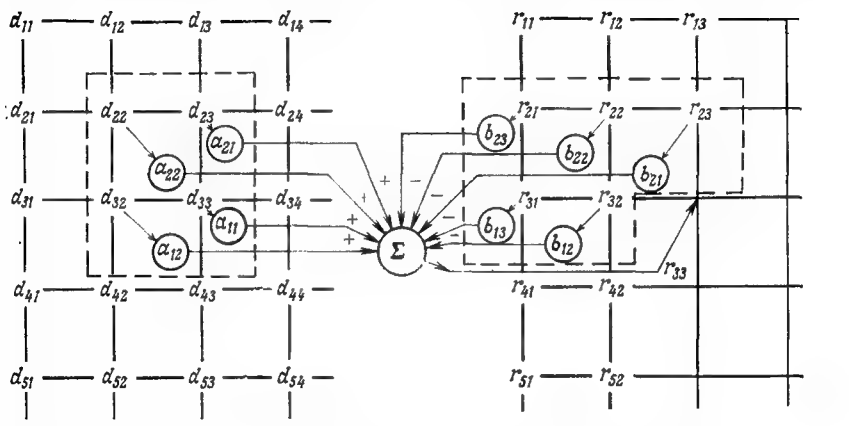


Фиг. 4.32. Реализация двумерного рекурсивного фильтра с использованием промежуточного массива переменных состояния.

Следует обратить внимание на то, что массив переменных состояния обеспечивает связь между входными и выходными массивами. Каждое новое состояние определяется только суммой входного отсчета и предыдущих состояний. Точно так же каждое новое выходное значение определяется только переменными состояния. Реализацию этого алгоритма иллюстрирует фиг. 4.32.

Замечания по реализации

Фиг. 4.32 может навести на мысль о том, что введенный дополнительный двумерный массив требует большего объема памяти ЦВМ, чем алгоритм фиг. 4.33, являющийся прямой реализацией (4.3). Это соответствовало бы истине, если бы все массивы, используемые при вычислениях, хранились в главном



Фиг. 4.33. Прямая реализация двумерного рекурсивного фильтра.

запоминающем устройстве (ЗУ) машины. Однако в случае обработки больших двумерных массивов это очень часто оказывается непрактичным. Более обычная ситуация при использовании больших массивов данных — хранение двумерных данных во внешнем накопителе большой емкости, откуда они переписываются в память ЦВМ для обработки одного столбца или одной строки за один раз. Чтобы проиллюстрировать сказанное, рассмотрим случай переноса данных по одному столбцу. При этом допущении фиг. 4.32 показывает, что для получения одного столбца выходных данных следует в качестве обрабатываемого массива использовать один столбец входных данных и производить обновление массива переменных состояния данными этого столбца, как это следует из (4.60). Тогда столбец выходных данных вычисляется путем обработки массива W согласно (4.61). При использовании этого метода число столбцов массива W, которые должны храниться в главном ЗУ машины в какой-то момент времени, не превышает N_a или N_b. Фиг. 4.33 показывает, что в случае прямой реализации алгоритма в главном ЗУ требуется одновременно хранить $N_a + N_b$ столбцов данных. Когда главное ЗУ процессора имеет ограниченную емкость, этот фактор может оказаться существенным.

Литература

1. *Huang T. H.*, IEEE Trans. Audio Electroacoust., **AU-20**, 158 (1972).
2. *Bliss G. A.*, Algebraic Functions, Am. Math. Soc., New York, 1933.
3. *Anderson B. D. O., Jury E. D.*, IEEE Trans. Audio Electroacoust., **AU-21**, 366 (1973).
4. *Ansel H. G.*, IEEE Trans. Circuit Theory, **CT-11**, 214 (1964).

5. *Cohn A.*, Math. J., **14**, 110 (1922).
6. *Fujiwara M.*, Math. J., **24**, 160 (1926).
7. *Jury E. I.*, Theory and Application of the *z*-transform Method, J. Wiley, New York, 1964.
8. *Barnett S.*, Matrices in Control Theory, Van Nostrand-Reinhold, London, 1971.
9. *Justice J. H., Shanks J. L.*, IEEE Trans. Automatic Control **AC-18**, 284 (1973).
10. *Shanks J. L., Treitel S., Justice J. H.*, IEEE Trans. Audio Electroacoust, **AU-20**, 115 (1972).
11. *Åstrom K. J.*, Introduction to Stochastic Control Theory, Academic Press, New York, 1970.
12. *Jury E. I.*, IEEE Trans. Circuit Theory, **DT-11**, 292 (1964).
13. *Rader C. M., Gold B.*, Proc. IEEE, **55**, 149 (1967). [Имеется перевод: ТИИЭР, **55**, № 2, 19 (1967).]
14. *Golden R. M., Kaiser J. F.*, Bell Syst. Tech. J., **43**, 1533 (1964).
15. *Диткин В. А., Прудников А. П.*, Операционное исчисление по двум переменным и его приложения, М., Физматгиз, 1958.
16. *Brown R. G., Nilsson J. W.*, Introduction to Linear Systems Analysis, J. Wiley. New York, 1962, pp. 121—125.
17. *Shanks J. L.*, Geophysics, **32**, 33 (1967).
18. *Burrus C. S., Parks T. W.*, IEEE Trans. Audio Electroacoust, **AU-18**, 137 (1970).
19. *Parks T. W., Burrus C. S.*, Applications and extensions of Prony's method to parameter identification and digital filtering, presented at 5th Princeton Conf. Inform. Sci. Syst. (March 1971).
20. *Wiggins R. A.*, On factoring the correlations of discrete multivariable stochastic processes, MIT Sci. Rep. 9 of Contract AF 19(604), 7378, pp. 127—152, 1965.
21. *Fletcher R., Powell M. J. D.*, Computer J., **6**, 163 (1963).
22. System/360 Scientific Subroutine Package (360-CM-03X) Version III, Programmer's Manual, Docmt. H20-0205-3 IBM, Data Proc. Div., White Plains, New York, USA (1968).
23. *Ansell H. G.*, IEEE Trans. Circuit Theory, **CT-11**, 214 (1964).
24. *Robinson E. A.*, Statistical Communication and Detection, Hafner Publishing Company, New York, 1967, 173—174.
25. *Read R., Treitel S.*, IEEE Trans. Geoscience Electronics, **GE-11**, 153 and 205 (1973).
26. *Gold B., Rader C. M.*, Digital Processing of Signals, McGraw-Hill Book Co., New York, 1969. (Имеется перевод: *Голд Б., Рэйдер Ч.* Цифровая обработка сигналов. — М.: Советское радио, 1973.)
27. *Dudgeon D. E.*, Two-dimensional recursive filtering, Ph. D. dissertation, Dept. of Electr. Eng., MIT, 1974.

5. Улучшение и реставрация изображения

Б. Фриден

Цель накопления данных состоит в получении содержательной информации об изучаемом явлении. К сожалению, явление часто не удается наблюдать непосредственно, и поэтому получаемые данные могут оказаться, например, линейной суперпозицией нужных величин. Этот линейный и простейший вид перемешивания информации повсеместно распространен в физических науках, встречаясь в таких различных областях, как атмосферная физика и медицинская диагностика (см. перечень, приведенный в разделе 5.1). Работающие в этих областях сталкиваются с общей проблемой — как разделить «перемешанные» данные (т. е. как реставрировать и улучшить данные, освободить их от смазывания, произвести инверсию свертки).

Решение этой проблемы осложняется явлением случайного шума, который неизбежно добавляется к данным в самом процессе их получения. В определенных случаях шум обусловливается природой информационного канала (скажем, турбулентностью атмосферы) и вносится еще до детектирования. Понятно, что шум является единственным препятствием идеальной реставрации требуемых данных. Однако если получаемые данные ограничены по полосе частот, идеальная реставрация в принципе не исключается. Более того, при реставрации даже в присутствии шума часто удается добиться значительного (двукратного и более) расширения полосы частот данных. По-видимому, наибольший интерес читателей в этой главе вызовет объяснение этого важного и современного результата (см. раздел 5.12).

Перейдем теперь к количественному рассмотрению проблемы. Мы будем пользоваться языком формирования оптического изображения, однако читатель может легко заменить его языком своей области (см. сопоставление различных областей в разделе 5.1). Несмотря на то что оптические изображения обычно двумерны, ради упрощения будем применять одномерные обозначения. Это позволит избежать громоздкости двойных аргументов, индексов и т. п. и сконцентрировать внимание на наиболее характерных чертах методов реставрации.

При формировании оптического изображения неизвестное пространственное распределение энергетической яркости

(radiance) $o(x)$, называемое «объектом», создает изображение в виде распределения энергетической освещенности (irradiance) $i(x)$, которое представляется совокупностью данных $i(y_m)$, $m = 1, \ldots, M$. Требуется оценить $o(x)$ на дискретном подмножестве точек x_n, $n = 1, \ldots, N$. Обозначим эту оценку символом $\hat{o}(x_n)$, $n = 1, \ldots, N$, либо просто $\hat{o}(x)$ для краткости. Соотношение между изображением и объектом выражается линейной формой

$$i(y_m) = \int_{-X}^{X} dx\, o(x)\, s(y_m; x) + n(y_m), \quad m = 1, \ldots, M, \tag{5.1}$$

где s — функция рассеяния точки всей системы, формирующей изображение, n — шум (случайный или иного вида). Обратите внимание на конечные пределы интегрирования X. Они обусловлены тем, что любой оптический прибор имеет ограниченное поле зрения; вследствие виньетирования имеем

$$s(y_m; x) = 0 \quad \text{для всех} \quad |x| > X. \tag{5.2}$$

Поэтому, даже если объект $o(x)$ имеет бесконечную протяженность, данные $i(y_m)$ формируются только за счет света, исходящего от существенно ограниченной области объекта. Нас интересует оценка, или реставрация, $o(x)$ только в пределах этой области.

Предыдущее замечание представляет не только академический интерес. Ниже будет показано, что благодаря конечности пределов X в (5.1) в отсутствие шума n объект $o(x)$ может быть реставрирован идеально, а при наличии значительного шума — в полосе частот, превышающей полосу частот данных $i(y_m)$.

Для того чтобы облегчить нахождение $o(x)$ на дискретном подмножестве x_n, $n = 1, \ldots, N$, удобно [1] заменить интеграл в (5.1) аппроксимирующей суммой; в результате получаем

$$i(y_m) = \sum_{n=1}^{N} w_n o(x_n)\, s(y_m; x_n) + n(y_m), \quad m = 1, \ldots, M. \tag{5.3}$$

Числа $\{w_n\}$ представляют собой веса входных значений, обеспечивающие получение необходимой точности аппроксимирующей суммы и определяемые, например, по правилу Симпсона.

В настоящее время не существует единого оптимального метода оценки $o(x)$ общего вида. Растущая масса эмпирических результатов заставляет думать, что существует «наилучший» (в любом заранее определенном смысле) метод реставрации, который зависит от вида имеющейся в распоряжении априорной информации об объекте и шуме. Для выбора метода реставрации оказались важными такие виды предварительной информации,

как форма объекта (например, плавная или резкая форма, т. е. объект ступенчатого или импульсного типа) и статистические характеристики объекта и шума.

Разумеется, при выборе метода должны учитываться и более второстепенные факторы, такие, как объем обрабатываемых данных и допустимая стоимость использования ЦВМ. Часто эти факторы приходится учитывать при выборе оптимального метода наряду с «чистыми» соображениями точности. Это обусловлено тем, что в настоящее время большинство надежных и точных методов, обеспечивающих расширение полосы частот данных в присутствии шума, оказываются слишком медленными для применения с протяженными объектами. Это в свою очередь объясняется тем, что большинство методов было разработано специально для одномерных задач, в которых такие операции, как итеративная свертка и итеративная инверсия матриц, не требуют слишком длительных вычислений.

По этим причинам мы ограничимся изучением предложенных методов реставрации, причем главным образом тех, которые появились за последние 10 лет. Выбор же метода для конкретного применения остается за самим пользователем.

5.1. Перечень неоптических приложений и библиография

Большинство современных успехов в решении уравнений (5.1) и (5.3) достигнуто исследователями, работающими в неоптических областях. Проявив большую изобретательность (отмеченную во многих случаях докторскими степенями), эти исследователи разработали алгоритмы реставрации, которые отличаются один от другого не меньше, чем сами области исследования. Некоторые из этих методов, наиболее перспективные для решения задач реставрации изображения, описываются ниже.

Количество и многообразие физических и математических задач, которые моделируются уравнением (5.1), удивительно велико. Для примера укажем перечень некоторых из них, приведенный Туми [2]. Столь же велика номенклатура встречающихся задач (причем, естественно, на каждую область исследования приходится по крайней мере одна задача). Ниже приводится неполный список этих областей с указанием некоторых задач и одной-двух типичных работ:

1. Атмосферная физика; дистанционное исследование: Тумп [2], Ямамото и Танака [3].

2. Фурье-спектроскопия; преобразование Фурье интерферограмм: Ванасс и Сакаи [4].

3. Геофизика; оценка энергетического спектра по функции автокорреляции: Бург [5], Лакосс [6].

4. Медицинская диагностика; восстановление изображений по их проекциям: Гордон и Герман [7].

5. Численный анализ; инверсия уравнений Фредгольма первого рода: Филлипс [1].

6. Радиоастрономия; обострение характеристик антенн: Шелл [8], Биро [9].

7. Спектроскопия; инверсная свертка: Джанссон [10].

8. Статистика; оценка плотности вероятности по измеренным моментам: Джейнис [11], Рагг и Доусон [12].

Читатели, вероятно, могут сами продолжить этот список.

5.2. Математическое введение, обозначения и определения

Мы уже определили физические понятия: объект $o(x)$, его желаемую оценку $\hat{o}(x)$, изображение $i(y)$ и шум $n(y)$, использованные в (5.1). Дадим теперь некоторые математические определения.

Если дана функция f пространственной координаты x, то ее преобразование Фурье F по пространственной частоте (в радианах на единицу длины) ω определяется как

$$F(\omega) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} dx\, f(x) \exp(-j\omega x). \tag{5.4}$$

Иногда операцию преобразования Фурье обозначают ПФ. Энергетический спектр $\phi_f(\omega)$, соответствующий спектру F, определяется как среднее по ансамблю

$$\phi_f(\omega) \equiv \langle |F(\omega)|^2 \rangle. \tag{5.4a}$$

В общем функции пространственной координаты x будут обозначаться строчными буквами, например f, а соответствующие преобразования Фурье — прописными, например F. Таким образом, имеем пары ПФ: $o(x)$ и $O(\omega)$, $i(y)$ и $I(\omega)$, $n(y)$ и $N(\omega)$. Исключением (и, должно быть, единственным) из этого правила является ПФ функции рассеяния точки s, которое будет обозначаться как $\tau(\omega)$ в соответствии с обозначением О'Нейла [13].

Что касается векторов и матриц, то вектор, соответствующий $f(x_n)$, $n = 1, \ldots, N$, будет обозначаться полужирной буквой $\boldsymbol{F}$, а матрица, соответствующая $s(y_m; x_n)$, будет обозначаться $[S]$. Матрицы и преобразования Фурье никогда не используются совместно, так что путаница исключается.

Если в качестве ПФ используется (5.4), то обратное соотношение между f и F имеет вид

$$f(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} d\omega\, F(\omega) \exp(+j\omega x). \qquad (5.5)$$

Функция $f(x)$ называется «ограниченной по полосе частот», или просто «ограниченной по частоте», если ее преобразование Фурье (спектр) имеет резкую границу: $F(\omega) = 0$ при $|\omega| > \Omega$. При этом частота Ω называется граничной частотой, и бесконечные пределы в (5.5) можно заменить на $\pm\Omega$.

Наоборот, функция $F(\omega)$ может быть названа «ограниченной по полосе пространства», если существует такая координата X, что $f(x) = 0$ при $|x| > X$. На обычном языке это означает, что $f(x)$ имеет ограниченную пространственную протяженность. Так, из (5.1) и (5.2) видно, что $o(x)$ имеет ограниченную пространственную протяженность, и, следовательно, $O(\omega)$ является ограниченной по полосе пространства.

При работе с функциями, ограниченными по полосе, полезно определить специальные функции Rect (x) и sinc (x):

$$\text{Rect}(x) \equiv \begin{cases} 1 & \text{при} \quad |x| \leqslant 1, \\ 0 & \text{при} \quad |x| > 1, \end{cases} \qquad (5.6а)$$

$$\text{sinc}(x) \equiv x^{-1} \sin(x). \qquad (5.6б)$$

Между понятиями (5.4), (5.6а) и (5.6б) имеется связь, обусловленная тем, что ПФ от Rect(x) выражается в виде 2 sinc(ω). Эта запись демонстрирует главное преимущество определений (5.4)—(5.6), которое состоит в устранении множителя 2π в аргументе функций.

Для ограниченной по частоте функции $f(x)$ выполняется теорема отсчетов Уиттекера — Шеннона[1])

$$f(x) = \sum_{n=-\infty}^{\infty} f(n\pi/\Omega)\, \text{sinc}(\Omega x - n\pi) \quad \text{при всех } x \qquad (5.7а)$$

и безымянная, но полезная дополнительная теорема

$$F(\omega) = (\pi/\Omega) \sum_{n=-\infty}^{\infty} f(n\pi/\Omega) \exp(-jn\pi\omega/\Omega) \quad \text{при} \quad |\omega| \leqslant \Omega. \qquad (5.7б)$$

Выражение (5.7а) иногда называют формулой идеальной интерполяции: если f известна на подмножестве точек x с интервалом π/Ω, то она известна всюду между этими точками. При

[1]) Для теории связи эта теорема была впервые доказана В. А. Котельниковым в 1933 г. — *Прим. перев.*

этом точки $x_n = n\pi/\Omega$ приобретают смысл узлов интерполяции или независимых степеней свободы функции f. Интервал π/Ω часто называют интервалом Найквиста.

Теоремы (5.7) можно доказать, используя (5.5) с конечными пределами интегрирования и замечая, что $F(\omega)$ отличается от нуля лишь на конечном интервале $|\omega| \leqslant \Omega$, что позволяет представить ее на этом интервале рядом Фурье. Из (5.5) следует, что коэффициенты этого ряда точно равны $f(n\pi/\Omega)$. Подстановка их в ряд Фурье дает (5.7б). Затем подставляем (5.7б) в (5.5) и после почленного интегрирования получаем (5.7а).

Если $O(\omega)$ полностью известна в конечной частотной области $|\omega| \leqslant \Omega$, то можно образовать оценку «главного значения», или «идеальную частотно-ограниченную» оценку объекта $\hat{o}_{\text{огр}}(x)$, в виде

$$\hat{o}_{\text{огр}}(x) = (2\pi)^{-1/2} \int_{-\Omega}^{\Omega} d\omega\, O(\omega) \exp(+j\omega x). \tag{5.8}$$

Поскольку идеальная частотно-ограниченная оценка имеет полностью детерминированный характер и прозрачный смысл, она может служить основой для сравнения с другими методами реставрации.

Далее мы будем часто пользоваться математическим тождеством, называемым теоремой Винера — Хинчина:

$$\int_{-\infty}^{\infty} dx\, f(x)\, g(x) \exp(-j\omega x) = \int_{-\infty}^{\infty} d\omega' F(\omega')\, G(\omega - \omega'). \tag{5.9}$$

Она легко выводится путем замены f и g в (5.9) на их преобразования Фурье (5.5) и интегрирования по x. Тождество

$$\int_{-\infty}^{\infty} dx \exp(\pm j\omega x) = 2\pi\delta(\omega) \tag{5.9а}$$

помогает выполнению этой операции. Здесь $\delta(\omega)$ — дельта-функция Дирака, определяемая как

$$\delta(x) = 0 \quad \text{при} \quad x \neq 0,$$

причем

$$\int_{-\infty}^{\infty} dx\, \delta(x) = 1. \tag{5.10}$$

Правая часть (5.9) определяет математическую операцию, называемую «сверткой». Для нее полезно ввести специальное обозначение

$$F \otimes G. \tag{5.11}$$

Наконец, полезно упомянуть «выделяющее» свойство дельта-функции Дирака. Когда $\delta(x)$ участвует в свертке, воспользовавшись свойством (5.10), можно получить выражение

$$\int_{-\infty}^{\infty} dx' f(x')\,\delta(x-x') = f(x). \tag{5.12}$$

Это означает, что единственное значение $f(x)$ выделяется из бесконечного множества значений $f(x')$, интегрируемых в левой части равенства. Это свойство дельта-функции наряду с (5.9) является наиболее употребительным на практике.

5.3. Интуитивные методы реставрации

Имеются по меньшей мере два метода нахождения $o(x)$, которые подсказывает наша интуиция. В *дискретной* задаче (5.3) функция $o(x)$ рассматривается как множество неизвестных $o(x_n)$ в системе линейных уравнений. Решение *непрерывной* задачи (5.1) для людей с инженерной подготовкой сводится к выполнению преобразования Фурье с последующей фильтрацией результата. Мы обсудим оба этих подхода (в п. 5.3.1 и разделе 5.8 соответственно) и посмотрим, что они могут дать.

5.3.1. Инверсия матрицы уравнения дискретного отображения

Уравнение (5.3) фактически представляет систему M уравнений, линейных относительно N неизвестных $o(x_n)$. Определив матрицу $[S] = \{w_n s(y_m; x_n)\}$ и векторы $\boldsymbol{I} = \{i(y_m)\}$, $\boldsymbol{N} = \{n(y_m)\}$ и $\boldsymbol{O} = \{o(x_n)\}$, можем записать (5.3) в виде

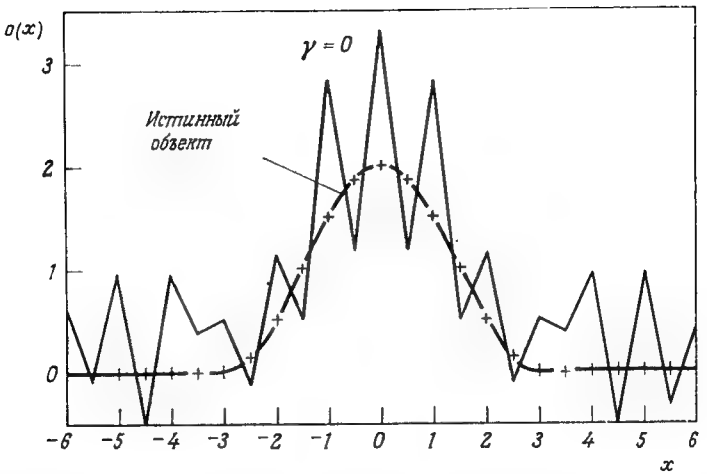
$$\boldsymbol{I} = [S]\,\boldsymbol{O} + \boldsymbol{N}. \tag{5.13}$$

При условии $M = N$ матрица $[S]$ станет квадратной матрицей, имеющей инверсию, и тогда $[O]$ находится как

$$\boldsymbol{O} = [S]^{-1}\boldsymbol{I} - [S]^{-1}\boldsymbol{N}. \tag{5.14}$$

Следовательно, в отсутствие шума (5.14) дает точное решение задачи.

Однако наличие даже слабого шума в данных $\boldsymbol{I}$ полностью разрушает решение, приводя к наложению на него паразитных осцилляций. Для иллюстрации этого факта служит фиг. 5.1 из работы Филлипса [1]. Максимальное значение шума составляет здесь 0,004 (среднее значение 0,0014) по шкале, где максимальный сигнал (без шума) достигает значения 3. Это дает чрезвычайно высокое и, по-видимому, недостижимое экспериментально значение отношения сигнал/шум $3/0{,}0014 = 2140$. Тем не менее



Фиг. 5.1. Иллюстрация искажений, присущих методу непосредственной инверсии матриц. Истинный объект соответствует гладкой кривой. Кривая с резкими изломами, помеченная индексом $\gamma = 0$, получена путем прямой инверсии согласно (5.14); при этом было принято $N = 0$. Основная причина отказа от данного метода состоит в том, что вопреки предположению здесь $N \neq 0$ и не предусмотрено никаких средств для сглаживания ошибок. В противоположность этому алгоритм Филлипса (5.22) обеспечивает сглаживание шума, зависящее от параметра γ. Точки на графике, обозначенные крестиками, получены с помощью алгоритма (5.22) при $\gamma = 0{,}01$ [1].

из фиг. 5.1 видно, что метод реставрации (5.14) дает очень плохой результат.

В чем причина такой гипертрофированной реакции данных на шум? В основном она состоит в том, что матрица $[S]$, выражающая функцию рассеяния точки, в значительной степени заполнена нулевыми и малыми элементами вблизи диагонали, а вследствие этого $[S]^{-1}$ содержит элементы с весьма большими значениями. Поэтому в точках y_m, где значения шума n_m конечны, член $[S]^{-1}\boldsymbol{N}$, выражающий ошибку в (5.14), становится весьма большим.

Причина того, что ошибка имеет осциллирующий характер, также объяснена Филлипсом [1]. Наблюдаемые $\{i_m\}$ инвариантны относительно синусоиды $\sin kx$ при больших k, которая добавляется к правильному решению $o(x)$. Инвариантность вытекает из аннулирования положительных и отрицательных составляющих вида $\sin kx$ при операции суперпозиции (5.3). Очевидно, что только осциллирующая ошибка, подобная $\sin kx$, может вызвать такое аннулирование. Поэтому при наличии в данных шума любая оценка $\hat{o}(x)$, которая удовлетворяет только (5.3), может оказаться искаженной из-за присутствия высокоча-

стотного осциллирующего члена. Для преодоления этой трудности требуется, очевидно, учет еще каких-то обстоятельств, помимо необходимости удовлетворять (5.3).

5.4. Метод сглаживания Филлипса

Из фиг. 5.1 можно сделать вывод, что осцилляция, привнесенная в $\hat{o}(x)$, влечет за собой появление больших значений производной $d^2o(x)/dx^2$, или (что эквивалентно) больших вторых разностей $\hat{o}_{n+1} - 2\hat{o}_n + \hat{o}_{n-1}$. Поэтому есть основания полагать, что решение $\hat{o}(x)$ с малыми вторыми разностями во всей области его определения окажется более точным. В то же время желательно, чтобы $\hat{o}(x)$ удовлетворяла (5.3) для входных данных. Поэтому имеет смысл поступиться согласованностью со входными данными ради обеспечения гладкости, для чего следует искать $\hat{o}(x)$, удовлетворяющее уравнению

$$\gamma \sum_{n=1}^{N} (\hat{o}_{n+1} - 2\hat{o}_n + \hat{o}_{n-1})^2 + \sum_{m=1}^{M} \hat{n}_m^2 = \min. \qquad (5.15)$$

Член $\hat{n}_m$ выражает «несогласованность» или «несовместность» левой и правой частей (5.3):

$$\hat{n}_m = i(y_m) - \sum_{n=1}^{N} w_n \hat{o}(x_n)\, s(y_m;\ x_n). \qquad (5.16)$$

Величина выбранного параметра γ определяет степень преобладания гладкости, выражаемой первым членом (5.15), над согласованностью.

5.4.1. Вывод

Решение $\hat{o}(x)$ задачи (5.15), (5.16) находится с помощью обычных методов вычисления. Подстановка частных производных $\partial/\partial\hat{n}_m$ в (5.15) и приравнивание нулю дает

$$\gamma \sum_{n=1}^{N} (\hat{o}_{n+1} - 2\hat{o}_n + \hat{o}_{n-1})(\partial\hat{o}_{n+1}/\partial\hat{n}_m - 2\,\partial\hat{o}_n/\partial\hat{n}_m + \partial\hat{o}_{n-1}/\partial\hat{n}_m) + \\ + \hat{n}_m = 0. \qquad (5.17)$$

Но требуемые частные производные можно найти из (5.14). Полагая $M = N$, это уравнение можно представить в виде

$$\hat{o}_l = \sum_k \alpha_{lk}(i_k - n_k), \quad \text{причем} \quad [S]^{-1} \equiv \{\alpha_{lk}\}, \qquad (5.18)$$

так что

$$\partial\hat{o}_j/\partial\hat{n}_k = -\alpha_{jk}. \qquad (5.19)$$

Поэтому (5.17) принимает вид

$$-\gamma \sum_{n=1}^{N} (\hat{o}_{n+1} - 2\hat{o}_n + \hat{o}_{n-1})(\alpha_{n+1,\,m} - 2\alpha_{nm} + \alpha_{n-1,\,m}) + \hat{n}_m = 0. \quad (5.20)$$

Это уравнение линейно относительно $\hat{o}$ и может быть записано в виде

$$-\gamma [B] \boldsymbol{O} + \boldsymbol{N} = 0 \quad (5.21)$$

с использованием матрицы $[B]$, определенной в (5.20). Исключая $\boldsymbol{N}$ из системы уравнений (5.13) и (5.21), получаем оценку объекта

$$\hat{\boldsymbol{O}} = ([S] + \gamma [B])^{-1} \boldsymbol{I} \quad (5.22)$$

и шума

$$\hat{\boldsymbol{N}} = -\gamma [B] \hat{\boldsymbol{O}}. \quad (5.23)$$

5.4.2. Обсуждение

Для проверки решения полагаем $\gamma = 0$, что, согласно (5.15), приводит к устранению сглаживания. Это находит свое отражение в результатах (5.22) и (5.23) и проявляется в том, что (5.22) сводится к решению (5.14) на основе инверсии матрицы при $\boldsymbol{N} = 0$. Но более интересно, что с ростом γ в решении (5.22) относительный вес гладкости увеличивается, а согласованности уменьшается — именно так, как этого требует (5.15). При излишне больших γ $\hat{\boldsymbol{O}}$ оказывается даже более смазанным, чем исходные данные $\boldsymbol{I}$.

Очевидно, что пользователь должен выбирать значение γ в промежутке между этими предельными значениями. И здесь вступает в силу выражение (5.23), дающее оценку шума $\hat{\boldsymbol{N}}$. Пользователь обычно имеет некоторые сведения о шуме в изображении; как правило, с какой-то точностью бывает известна дисперсия шума. При этом корректное решение (5.23), т. е. решение, основанное на применении почти правильного значения γ, должно дать оценку шума $\hat{\boldsymbol{N}}$ с дисперсией, приблизительно соответствующей известной. Следовательно, чтобы найти правильное значение γ, пользователь должен получить несколько решений (5.22) и (5.23), основанных на различных значениях γ, и затем выбрать решение, которое дает оценку шума, наиболее хорошо согласующуюся с априорной информацией о шуме (например, до дисперсии).

Точки, отмеченные знаками «+» на фиг. 5.1, соответствуют использованию в формуле реставрации (5.22) значения $\gamma = 0{,}01$. Дальнейшее обсуждение вопроса о выборе γ можно найти у Филлипса [1].

5.4.3. Достоинства и недостатки метода в применении к улучшению изображения

Основное преимущество метода состоит в том, что он позволяет сглаживать шум без излишнего смазывания выхода (при условии правильного выбора γ) и в предельном случае отсутствия шума обеспечивает получение идеального выхода $\hat{o}(x)$.

Что касается времени вычислений, то (5.22) дает решение в замкнутой форме, которое требует выполнения лишь инверсии двух матриц и одного векторного умножения. В этом заключается определенное преимущество для решения одномерных задач.

Однако, переходя к проблеме улучшения двумерных изображений, мы замечаем, что матрицы $[S]$ и $[B]$ содержат по $N \times N$ элементов, где N — число точек входного изображения. Поэтому при входной матрице 32×32 указанные матрицы имеют размер 1024×1024! Инверсия матриц такого размера (если не касаться специальных случаев) представляет грандиозную, но не неразрешимую задачу, учитывая растущую производительность вычислительных машин.

С оптической точки зрения метод имеет еще некоторые недостатки:

1. Неопределенность выбора γ.

2. Неограниченный характер оценки. Хотя мы знаем, что $o(x)$ положительно, решение $\hat{o}(x)$ из (5.22) не обязательно является таковым. Ниже будет показано, что методы реставрации с ограничениями имеют важные преимущества по сравнению с методами, не учитывающими ограничений.

3. Осциллирующий шум в $\hat{o}(x)$ подавляется за счет сглаживания $\hat{o}(x)$ с использованием минимизации вторых разностей. Однако во многих оптических приложениях вторые разности интересующего объекта велики, как это имеет место в случае звездных полей, контуров и линейных оптических спектров. Поэтому для оптических целей следовало бы использовать более хороший критерий, чем минимум вторых разностей. Возможный подход к решению этой проблемы описывается ниже.

5.5. Обобщенный метод Туми

Мы могли заметить, что сумма вторых разностей по Филлипсу (5.15) является квадратичной относительно $\hat{o}(x)$. Пытаясь использовать иные формы ограничения, чем минимизация вторых разностей, Туми [2] обнаружил, что решения, подобные (5.22), могут быть получены на основе любого ограничения, квадратичного относительно $\hat{o}(x)$. В качестве примеров можно

указать минимизацию третьих разностей

$$\gamma \sum_{n=1}^{N} (\hat{o}_{n+1} - 3\hat{o}_n + 3\hat{o}_{n-1} - \hat{o}_{n-2})^2 + \sum_{m=1}^{M} \hat{n}_m^2 = \min, \quad (5.24)$$

минимизацию отклонения от известной кривой, т. е. от изображения $p(x)$,

$$\gamma \sum_{n=1}^{N} (\hat{o}_n - p_n)^2 + \sum_{m=1}^{M} \hat{n}_m^2 = \min, \quad (5.25)$$

или даже их линейную комбинацию. Для такого общего класса ограничений он нашел решение

$$\hat{\boldsymbol{O}} = ([S]^T [S] + \gamma [H])^{-1} ([S]^T \boldsymbol{I} + \gamma \boldsymbol{P}). \quad (5.26)$$

Вывод этой формулы аналогичен выводу (5.22), за исключением того, что новые формы (5.24), (5.25) и им подобные дифференцируются по $\hat{o}_n$, а не по $\hat{n}_m$.

В (5.26) форма матрицы H зависит от конкретных используемых ограничений, а $\boldsymbol{P} = 0$, если «смещающая» кривая $p(x)$ неизвестна. Например, при использовании критерия Филлипса (5.15) H имеет вид

$$\begin{bmatrix} 1 & -2 & 1 & 0 & 0 & — & — & — \\ -2 & 5 & -4 & 1 & 0 & — & — & — \\ 0 & 1 & -4 & 6 & -4 & 1 & — & — \\ 0 & 0 & 1 & -4 & 6 & -4 & 1 & — \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}, \quad (5.27)$$

тогда как при минимизации третьих разностей H, согласно [14], принимает вид

$$\begin{bmatrix} 1 & -3 & 3 & -1 & 0 & 0 & 0 & — & — \\ -3 & 10 & -12 & 6 & -1 & 0 & 0 & — & — \\ 3 & -12 & 19 & -15 & 6 & -1 & 0 & — & — \\ -1 & 6 & -15 & 20 & -15 & 6 & -1 & — & — \\ 0 & -1 & 6 & -15 & 20 & -15 & 6 & -1 & — \\ 0 & 0 & -1 & 6 & -15 & 20 & -15 & 6 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}. \quad (5.28)$$

Другой полезный аспект решения (5.26) состоит в том, что матрица $[S]$ может не быть квадратной, как это требовалось в решении Филлипса (5.22). Тем самым обеспечивается решение задач как с излишним, так и с недостаточным ограничением, в которых либо количество входных данных (точек изображения)

превышает количество неизвестных значений объекта, либо имеется обратное соотношение. В случае излишнего ограничения имеется, например, возможность выполнить полезное усреднение данных в (5.26). Туми [14] приводит примеры приложения этого метода в форме таблицы сравнения $\hat{o}(x)$ с другими оценками.

5.5.1. Достоинства и недостатки метода в применении к улучшению изображений

В целом метод Туми имеет те же достоинства и недостатки, что и метод Филлипса, но обладает следующими дополнительными преимуществами:

1. Возможность решения задач с излишним или недостаточным ограничением.

2. Гибкость в выборе критерия сглаживания. В то же время это свойство может рассматриваться как недостаток, поскольку пользователь теперь вынужден принимать больше решений, тогда как прежде он должен был просто выбрать γ.

5.6. Последующие методы

В последующих работах, посвященных решению уравнения дискретного отображения (5.3), исходили из предположения, что пользователь располагает определенной статистической информацией относительно $o(x_n)$ и(или) $n(y_m)$; в частности, это могут быть ковариационные матрицы (см., например, работу Стрэнда и Вестуотера [15]). Однако обычно информации такого рода не имеется в распоряжении тех, кто занимается проблемой улучшения двумерных изображений, и поэтому мы не будем проводить дальнейшее изучение подобных методов.

5.7. Линейные или нелинейные методы?

Обсудив линейные методы инверсии *дискретного* (матричного) варианта отображения согласно (5.3), обратимся теперь к задаче линейной инверсии (или фильтрации) *непрерывного* варианта отображения (5.1). Этой задаче посвящено огромное количество исследований. При желании читатели могут обратиться к библиографическому указателю, составленному недавно Праттом [16], либо к перечню литературы в работе Хуанга и др. [17].

Однако кроме затруднений, вызываемых изобилием работ, есть более веская причина не принимать точку зрения об универсальности линейных методов. Выражаясь просто, существуют более хорошие методы обработки, чем линейные. Под более хорошими мы понимаем методы, обеспечивающие повышенную

точность выхода $\hat{o}(x)$, а совсем не увеличенную скорость вычислений. К таким методам относятся схемы экстраполяции полосы частот, разработанные Бургом [5], Биро [9], Шеллом [8], Джанссоном [10] и Фриденом [18, 19]. Все они являются нелинейными и, за исключением метода Бурга, итеративными.

Продолжающееся использование методов фильтрации оправдывается тем, что они обеспечивают высокую скорость вычислений и применяются в тех случаях, когда фильтровая оценка уже имеет достаточную полосу частот. Некоторым источникам, образующим смазанное изображение (в частности, в результате движения), присуща бесконечная полоса частот, и поэтому они вообще не требуют экстраполяции. Кроме того, у объектов, имеющих в основном сглаженный характер, спектр сосредоточен около начала координат частотной области (соответствующего постоянной составляющей).

Что касается скорости, то появившийся недавно алгоритм БПФ (быстрого преобразования Фурье), разработанный Гудом [20] и Кули и Тьюки [21], обеспечил применение методов фильтрации к двумерным изображениям большого размера (порядка 64×64 элемента). После доработки Синглтоном [22] этот алгоритм в принципе стал пригоден для осуществления быстрого преобразования Фурье над точечной решеткой любого размера. Опубликованы многочисленные примеры применения метода фильтрации (см., например, работу Мак-Гламери [23]).

Менее важным преимуществом является легкость аналитического исследования методов линейной фильтрации (в отношении предельного разрешения, передачи шума и т. п.). В противоположность этому для анализа нелинейных методов приходится применять метод проб и ошибок.

Кроме того, можно привести доводы, подтверждающие, что иногда линейные методы могут быть оптимальными, но, как будет показано ниже, для оптического случая это неверно.

Если в каждой точке y_m изображение (т. е. сигнал, получаемый в результате свертки) и шум в (5.1) образуют совместное нормальное распределение, то наилучшую среднеквадратичную оценку сигнала дает линейная комбинация из $i(y_m)$ [24, стр. 256]. С другой стороны, если сигнал и шум образуют независимые нормальные распределения, то оценка сигнала по максимуму правдоподобия также будет линейной [25, стр. 417]. Поскольку объект $o(x)$ линейно связан с изображением (через детерминированную функцию рассеивания), то для указанных двух случаев его оценка также будет линейной комбинацией данных $i(y_m)$.

Однако особенность оптических изображений состоит в том, что они не могут иметь статистических характеристик нормального типа, поскольку значения изображения могут быть только

положительными. Поэтому основная предпосылка приведенных соображений в пользу оптимальности линейной оценки не имеет силы в оптическом случае. С другой стороны, это согласуется с эмпирическим фактом, что нелинейные методы (перечисленные выше) обладают явными преимуществами. Более того, нелинейность в этих методах нужна именно для того, чтобы наложить особое ограничение на получаемый выход, который должен быть положительным (и отличным от нормального).

5.8. Принцип метода фильтрации

Если функция рассеяния точки $s(y_m; x)$ в (5.1) инвариантна относительно сдвига (или, как иногда говорят, выполняется условие стационарности или изопланатизма), то $s(y_m; x) = s(y_m - x)$ и преобразование Фурье от (5.1) дает

$$I(\omega) = O(\omega)\,\tau(\omega) + N(\omega) \tag{5.29}$$

(см. обозначения в разделе 5.2).

Метод фильтрации состоит в умножении спектра $I(\omega)$ на выбранную функцию фильтрации $Y(\omega)$:

$$\hat{O}(\omega) \equiv Y(\omega)\,I(\omega) = \tag{5.30а}$$

$$= Y(\omega)\,\tau(\omega)\,O(\omega) + Y(\omega)\,N(\omega) \tag{5.30б}$$

с последующим преобразованием Фурье, дающим выход в виде

$$\hat{o}(x) = (2\pi)^{-1/2} \int\limits_{-\Omega_p}^{\Omega_p} d\omega\,\hat{O}(\omega)\,e^{i\omega x}. \tag{5.30в}$$

Здесь Ω_p — граничная частота, которая по ряду причин, обсуждаемых ниже, должна быть конечной.

С другой стороны, такая операция фильтрации может быть полностью выполнена путем свертки в пространственной (исходной) области, поскольку (5.30а) и (5.30в) с учетом тождества (5.9) эквивалентны единой операции

$$\hat{o}(x) = \int\limits_{-\infty}^{\infty} dx' i(x')\,y(x - x'). \tag{5.31}$$

Наконец, для целей последующего анализа полезно определить полную функцию передачи

$$\tau_{\text{полн}}(\omega) \equiv \tau(\omega) \cdot Y(\omega), \tag{5.32}$$

описывающую совместно операции смазывания изображения и последующей фильтрации. Подстановка (5.32) в (5.30б) показывает, что сигнальная часть выхода $\hat{o}(x)$ связана с истинным

$o(x)$ через полную функцию рассеяния $s_{\text{полн}}(x)$, включающую в себя как смазывание изображения, так и последующую свертку:

$$\hat{o}_{\text{сиг}}(x) = \int_{-\infty}^{\infty} dx' o(x') s_{\text{полн}}(x - x'), \tag{5.33}$$

где

$$s_{\text{полн}}(x) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} d\omega \tau(\omega) Y(\omega) e^{j\omega x}.$$

Каждый из методов линейной фильтрации можно проанализировать в отношении разрешения и величины боковых лепестков, исследовав $s_{\text{полн}}(x)$; поэтому мы будем определять эту функцию для каждого из рассматриваемых методов.

5.9. Инверсная фильтрация

Уравнение (5.29) может быть формально решено относительно $O(\omega)$ в виде

$$O(\omega) = I(\omega)/\tau(\omega) - N(\omega)/\tau(\omega) \quad \text{при} \quad \tau(\omega) \neq 0. \tag{5.34}$$

Очевидно, что на частотах, где τ конечно, а отношение N/I достаточно мало, аппроксимация выражения (5.34) в виде

$$\hat{O}(\omega) = I(\omega)/\tau(\omega), \quad \tau(\omega) \neq 0 \tag{5.35}$$

обеспечивает точную реставрацию $\hat{O}$. В этом и состоит метод инверсной фильтрации, называемый иногда также методом компенсации функции передачи или частотной характеристики. Если воспользоваться терминологией раздела 5.8, то функция фильтрации

$$Y(\omega) = \text{Rect}(\omega/\Omega_p)\tau(\omega)^{-1}, \tag{5.36}$$

используемая в этом методе, служит для получения полной функции передачи

$$\tau_{\text{полн}}(\omega) = \text{Rect}(\omega/\Omega_p) \tag{5.37}$$

и полной функции рассеяния

$$s_{\text{полн}}(x) = \text{sinc}(\Omega_p x). \tag{5.38}$$

Величина $2\Omega_p$ представляет собой полосу частот, в которой производится обработка методом фильтрации. Из (5.37) и (5.33) с очевидностью следует, что Ω_p желательно иметь как можно

большим; в пределе при $\Omega_p \to \infty$ $s_{\text{поли}}(x)$ стремится к $\delta(x)$. Однако свойство оптического изображения таково, что

$$\tau(\omega) = 0 \quad \text{для всех} \quad |\omega| \geqslant \Omega, \tag{5.39}$$

где Ω пропорциональна наименьшей оптической апертуре в системе отображения. Поэтому $Y(\omega)$ в (5.36) оказывается неопределенным, если не соблюдается условие $\Omega_p < \Omega$.

Таким образом, при инверсной фильтрации качество сигнала ограничивается полной ФРТ (5.38), для которой расстояние разрешения x_0 (т. е. координата первого нуля по оси x) выражается как

$$x_0 = \pi/\Omega_p, \tag{5.40}$$

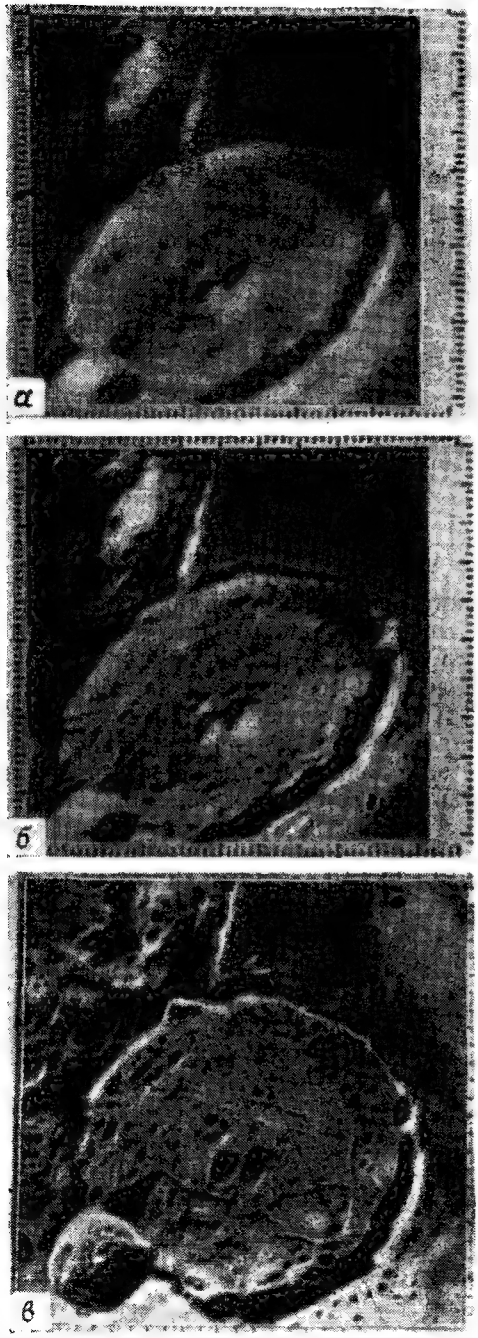
а наибольший боковой лепесток составляет около $2(3\pi)^{-1} \approx 0{,}22$. Однако здесь следует подчеркнуть, что (5.40) *не есть наименьшее* x_0 для любых фильтров с полосой пропускания $2\Omega_p$. Далее это будет четко показано на примерах. Рассмотрение других проблем, связанных с шумом, отложим пока на более поздний срок. Наконец, следует упомянуть проблему изолированных нулей $\tau(\omega)$ в пределах обрабатываемой полосы частот; ее обсуждение см., например, в работе Сондхи [26].

Несмотря на все эти проблемы, именно инверсная фильтрация обеспечила получение самых впечатляющих примеров улучшения больших изображений. Пример улучшения приведен на фиг. 5.2. Своему широкому применению инверсная фильтрация обязана главным образом использованию алгоритма быстрого преобразования Фурье (раздел 5.7), который благодаря высокой скорости позволяет осуществлять преобразование Фурье над большими двумерными массивами отсчетов изображения. С другой стороны, инверсная фильтрация имеет следующие недостатки:

1. Разрешение (5.40) часто оказывается недостаточным для практического применения. Это следует из того, что функция (5.37), выбранная в качестве полной функции передачи, не оптимальна по разрешению. Другие возможности обсуждаются в п. 5.10.1 в связи с выражениями (5.83) и (5.84).

2. Боковые лепестки величиной 0,22 часто оказываются неприемлемо большими из-за ложных деталей, которые они порождают в $\hat{o}(x)$.

3. Нефизические (отрицательные) $\hat{o}(x)$ часто охватывают протяженные области на оси x. Это связано с предыдущим недостатком, поскольку положительно-ограниченный выход не смог бы вызвать появления больших отрицательных, а тем более больших положительных выбросов в тех областях, где объект имеет нулевое излучение. Наличие паразитных положительных



Фиг. 5.2. *а* — исходное изображение лунного кратера Гассенди, полученное с помощью 60-сантиметрового телескопа фирмы Aerospace Corporation в обсерватории Сан Фернандо. Смазывание обусловлено турбулентностью атмосферы. *б* — улучшенный вариант этого изображения с использованием инверсной фильтрации. *в* — «истинное» изображение кратера Гассенди, полученное с борта автоматической станции «Лунар Орбитер». Как видно из представленных фотографий, многие детали, имеющиеся в изображении *в* и отсутствующие в *а*, были восстановлены в *б*. Длительная экспозиция при атмосферной турбулентности характеризуется приблизительно гауссовой функцией передачи, которая не ограничена по полосе частот и поэтому весьма эффективно компенсируется с помощью инверсной фильтрации [2].

выбросов привело бы к появлению соседних отрицательных выбросов, чтобы скомпенсировать их вклад в интеграл изображения (5.1) и обеспечить нулевое излучение. (См. обсуждение этого вопроса в конце п. 5.3.1.)

5.9.1. Оптимальная полоса частот обработки

Мы уже видели, что, исходя только из разрешения (5.40), желательно иметь Ω_p как можно большей. С другой стороны, поскольку в оптических системах происходит ослабление $\tau(\omega)$ с ростом ω, т. е.

$$\tau(\omega) \to 0 \quad \text{при} \quad \omega \to \Omega_p < \Omega, \tag{5.41}$$

алгоритм (5.35) приобретает член ошибки $N(\omega)/\tau(\omega)$, который неограниченно растет, когда $\omega \to \Omega_p$. Очевидно, имеется возможность оптимизации выбора Ω_p путем выравнивания ошибок выхода $\hat{o}(x)$, вызываемых недостаточным разрешением и возрастанием шума. Оптимизация может быть выполнена следующим образом. (Этот вывод не встречается в открытой литературе, поэтому мы приводим его полностью; кроме того, он прекрасно демонстрирует возможности анализа, которые предоставляют линейные методы.)

При любом x ошибка $e(x)$ в $\hat{o}(x)$ из-за применения фильтра (5.36) в обратном преобразовании Фурье, учитывая (5.29), составляет

$$\begin{aligned} e(x) &\equiv o(x) - \hat{o}(x) = \\ &= \int_{-\infty}^{\infty} d\omega\, O(\omega)\, e^{i\omega x} - \int_{-\Omega_p}^{\Omega_p} d\omega \left[O(\omega) + N(\omega)/\tau(\omega)\right] e^{i\omega x} = \\ &= \left(\int_{-\infty}^{-\Omega_p} + \int_{\Omega_p}^{\infty}\right) d\omega\, O(\omega) e^{i\omega x} - \int_{-\Omega_p}^{\Omega_p} d\omega \left[N(\omega)/\tau(\omega)\right] e^{i\omega x}. \end{aligned} \tag{5.42}$$

Здесь коэффициенты $(2\pi)^{-1/2}$ отброшены.

Будем считать, что оптимальная Ω_p должна минимизировать среднеквадратичную ошибку

$$\varepsilon^2 \equiv \left\langle \int_{-\infty}^{\infty} dx\, |e(x)|^2 \right\rangle. \tag{5.43}$$

Комбинируя (5.42) и (5.43) и воспользовавшись еще определением (5.4а), получаем

$$\varepsilon^2 = 2 \int_{\Omega_p}^{\infty} d\omega\, \phi_0(\omega) + 2 \int_{0}^{\Omega_p} d\omega\, \phi_n(\omega)/|\tau(\omega)|^2, \tag{5.44}$$

где учтены также предположения о статистических характеристиках:

$$\langle N(\omega)^* O(\omega')\rangle = 0 \text{ (объект и шум некоррелированы)},$$
$$\langle O(\omega)^* O(\omega')\rangle = \phi_0(\omega)\,\delta(\omega-\omega') \text{ (некоррелированный объект)}, \quad (5.45)$$
$$\langle N(\omega)^* N(\omega')\rangle = \phi_n(\omega)\,\delta(\omega-\omega') \text{ (некоррелированный шум)},$$

и предположение о симметрии функций, выражающих энергетические спектры ϕ_0 и ϕ_n, а также функции передачи τ.

Мы хотим минимизировать ε^2 в (5.44), выбирая Ω_p. В соответствии с этим запишем

$$\partial(\varepsilon^2)/\partial\Omega_p = 0, \quad (5.46)$$

откуда получается условие

$$-\phi_0(\Omega_p) + \phi_n(\Omega_p)|\tau(\Omega_p)|^2 = 0, \quad (5.47)$$

или

$$|\tau(\Omega_p)|^2 = \phi_n(\Omega_p)/\phi_0(\Omega_p). \quad (5.48)$$

Таким образом, получено довольно интересное соотношение — оптимальная полоса частот обработки определяется частотой, для которой модуль функции передачи равен корню из отношения сигнал/шум. В частном случае дифракционно-ограниченной оптики, когда

$$\tau(\omega) = \begin{cases} 1-|\omega|/\Omega & \text{при } |\omega| \leqslant \Omega, \\ 0 & \text{при } |\omega| > \Omega, \end{cases} \quad (5.49)$$

и при условии, что отношение сигнал/шум не зависит от частоты, (5.48) позволяет получить решение для Ω_p в явном виде:

$$\Omega_p = \Omega[1-(\phi_n/\phi_0)^{1/2}]. \quad (5.50)$$

Следовательно, оптимальная полоса частот обработки сокращается от полной оптической полосы до полосы, в которой корень из отношения сигнал/шум отличен от нуля. Так, если шум отсутствует, то из (5.50) следует, что существует инверсный фильтр для всей полосы частот; если же отношение сигнал/шум равно единице, то из (5.50) следует, что инверсный фильтр вообще не существует!

5.9.2. Использование функций окна

Для объектов общего вида по мере подхода к граничным значениям частоты $\pm\Omega_p$ спектр $\hat{O}(\omega)$ в (5.35) не приближается плавно к нулю. Возникающий резкий переход от конечного $\hat{O}(\omega)$ к нулю в интеграле Фурье (5.30в) вызывает появление в $\hat{o}(x)$ сильно флуктуирующих или ложных деталей. Например, если

истинный $o(x) = \delta(x)$, то $\hat{O}(\omega) = 1$; однако (5.36) вызывает резкую отсечку спектра у соответствующей граничной частоты, что приводит к реставрации $\hat{o}(x)$ в виде функции $\mathrm{sinc}(\Omega_p x)$, отрицательные выбросы которой достигают уровня 0,22.

Общий метод устранения этого эффекта состоит во введении дополнительной чисто математической функции $W(\omega)$, на которую умножается $Y(\omega)$ в (5.36) с тем, чтобы обеспечить плавный подход к нулю $\tau_{\text{полн}}(\omega)$ в (5.37) при $\omega \to \Omega_p$. В результате получается новый выход $\hat{o}_W(x)$, выражаемый как

$$\hat{o}_W(x) = (2\pi)^{-1/2} \int_{-\Omega_p}^{\Omega_p} d\omega\, W(\omega)\, \hat{O}(\omega) \exp(+j\omega x) \qquad (5.51)$$

(см., например, [27]). Роль функции $W(\omega)$ состоит в том, чтобы сделать произведение $W \cdot \hat{O}$ близким к $\hat{O}$, когда $\hat{O}$ конечно, и плавно приближающимся к нулю при $\omega \to \Omega_p$. Для этой цели с определенным успехом были использованы такие специализированные функции окна, как функция эллиптической дуги $W(\omega) = (1 - \omega^2/\Omega_p^2)^{1/2}$, фильтр Ханнинга $W(\omega) = \frac{1}{2}[1 + \cos(\pi\omega/\Omega_p)]$ (см. [28]) и т. п. Вывод оптимальной функции окна, основанный на ограничении резкости, приводится в п. 5.11.1.

Несомненно, подобные функции выполняют задачу сглаживания высокочастотных деталей в $\hat{o}(x)$. Чтобы показать это, заметим, что новая форма реставрации $\hat{o}_W(x)$ связана с прежней (т. е. с формой чистой инверсной фильтрации) соотношением

$$\hat{o}_W(x) = \int_{-\infty}^{\infty} dx' \hat{o}(x')\, w(x - x'). \qquad (5.52)$$

Отсюда следует, что при любой положительной $w(x)$ оценка $\hat{o}_W(x)$ должна быть более гладкой, чем $\hat{o}(x)$.

Недостаток описанного метода состоит в том, что отсутствует возможность различать существенные и привнесенные части $\hat{o}(x)$ и поэтому «вслепую» сглаживаются как те, так и другие, что с очевидностью следует из (5.52). Таким образом, за достигаемый положительный эффект здесь приходится расплачиваться общим сглаживанием всего изображения, которое, впрочем, в определенных случаях оказывается приемлемым.

С использованием функции окна полная ФРТ, связывающая $\hat{o}_W(x)$ с истинным объектом $o(x)$, принимает вид

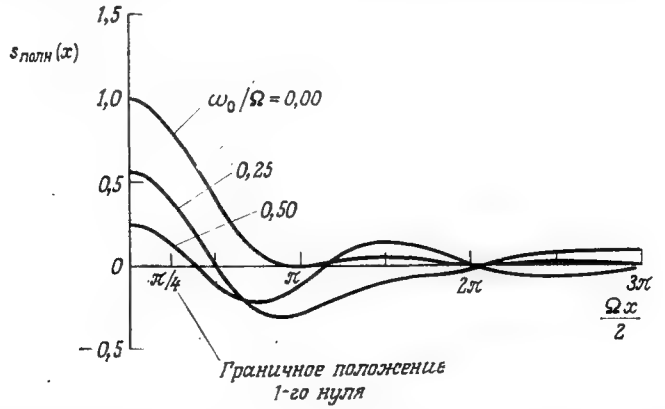
$$s_{\text{полн}}(x) = (2\pi)^{-1/2} \int_{-\Omega_p}^{\Omega_p} d\omega\, W(\omega)\, e^{j\omega x}, \qquad (5.53)$$

измененный по сравнению с (5.38) для случая простой инверсной фильтрации. Зависимость $s_{\text{полн}}(x)$ в (5.53) от $W(\omega)$ точно такая же, как зависимость интенсивности дифракционной картины от функции зрачка в оптике. Это позволяет использовать в качестве функции окна все прямоугольные аподизаторы Жакино и Досье [29]. [Цель применения этих аподизаторов состоит в подавлении боковых лепестков в дифракционной картине точки, т. е. та же, что и в нашем случае. За достигаемый эффект приходится расплачиваться уменьшением разрешения в $s_{\text{полн}}(x)$.]

С другой стороны, подобного рода оптическое соответствие позволяет вместо $W(\omega)$ использовать в (5.53) также функции «экстраполяции зрачка» Фридена [30], которые вводились с целью обеспечить сверхразрешение (задача, противоположная аподизации). Эта возможность обсуждается в п. 5.13.1. Еще одна функция $W(\omega)$, позволяющая улучшить разрешение, определяется в п. 5.10.1.

5.10. Прочие частные линейные методы

Под частным линейным методом мы понимаем любой метод, который в отсутствие шумов в данных $i(y_m)$ позволяет получить $\hat{o}(x)$, более похожий на $o(x)$, чем необработанное изображение



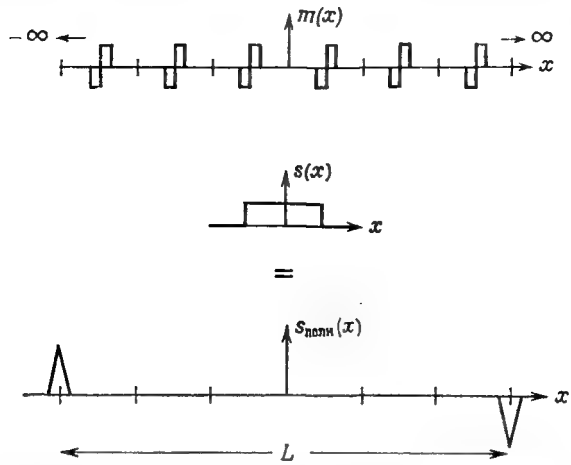
Фиг. 5.3. Функция рассеяния $s_{\text{полн}}(x)$, получаемая в результате подавления в изображении полосы частот $|\omega| \leqslant \omega_0$. Характеристики системы отображения определялись дифракционно-ограниченной оптикой с граничной частотой Ω. Показаны $s_{\text{полн}}(x)$, получаемые при различных значениях ω_0/Ω. Из графиков видно, что по мере приближения ω_0 к Ω разрешение улучшается, но одновременно увеличиваются боковые лепестки. Подобная взаимосвязь характерна для всех линейных методов реставрации. (Автор выражает благодарность своему аспиранту Вагнеру за расчет приведенных кривых.)

$i(x)$. Эти методы не обладают способностью подавлять шум, как это свойственно, например, методу фильтрации по Винеру.

Существует множество интересных методов такого типа, и некоторые из них кратко описываются ниже.

5.10.1. Подавление постоянной составляющей

К сожалению, недостаточно широко известно, что простое приписывание нулевых значений спектру $I(\omega)$ в конечной полосе частот с центром в начале координат обеспечивает улучшение выхода (см. [31]). В изображении точки, обрабатываемом таким способом, первый нуль смещается вниз по частоте (например, асимптотически приближается к $^1/_4$ значения частоты, занимаемого нулем необработанного дифракционно-ограниченного изображения) по мере того, как «нулевая» полоса частот спектра расширяется в сторону Ω. Фиг. 5.3 иллюстрирует характер



Фиг. 5.4. Иллюстрация принципа действия противосверточной маски $m(x)$, которая позволяет уменьшить искажения из-за движения до любой желаемой величины. Свертка $m(x)$ с искажающей функцией $s(x)$ приводит к образованию двух узких импульсов, ширина которых определяется прямоугольной функцией в составе $m(x)$. Поэтому любое изображение, смазанное в результате прямолинейного движения, может быть превращено в негативно-позитивную пару улучшенных изображений объекта. Поскольку разнос импульсов L может быть сколь угодно велик и зависит только от довольно свободно выбираемой $m(x)$, два реставрируемых изображения можно всегда сделать неперекрывающимися, т. е. не создающими взаимных помех. В этом состоит преимущество данного метода перед методом реставрации с помощью дифференцирования, при котором два улучшенных изображения разносятся лишь на фиксированную часть смещения при движении [32].

смещения первого нуля в этой ситуации, которую можно описать функцией окна

$$W(\omega) = \begin{cases} 0 & \text{при} \quad |\omega| \leqslant \omega_0, \\ \tau(\omega) & \text{при} \quad \omega_0 \leqslant |\omega| \leqslant \Omega \end{cases} \tag{5.54}$$

и полной функцией рассеяния [1])

$$s_{\text{полн}}(x) = \left(\int_{-\Omega}^{-\omega_0} + \int_{\omega_0}^{\Omega} \right) d\omega\, \tau(\omega)\, e^{i\omega x} = s(x) - \operatorname{sinc}(\omega_0 x) \otimes s(x), \tag{5.55}$$

полученной с использованием тождества (5.9).

5.10.2. Устранение искажений движения путем свертки с маской

В случае прямолинейного движения искажение изображения выражается функцией рассеяния $s(x) = \operatorname{Rect}(x/L)$. Как показал Суинделл [32], свертка $s(x)$ с простой функцией маски $m(x)$, показанной на фиг. 5.4, обеспечивает получение полной функции рассеяния $s_{\text{полн}}(x)$, имеющей два узких пика произвольной ширины, расположенные на произвольном расстоянии. Покажем, что функция рассеяния $s_{\text{полн}}(x)$ связывает оценку $\hat{o}(x)$ с неизвестной функцией $o(x)$. Полагая, что s в (5.1) инвариантна относительно сдвига, и используя обозначение свертки (5.11) и определение $s_{\text{полн}}$, имеем

$$\hat{o}(x) \equiv m \otimes i = m \otimes [o \otimes s] = o \otimes [m \otimes s] \equiv o \otimes s_{\text{полн}}. \tag{5.56}$$

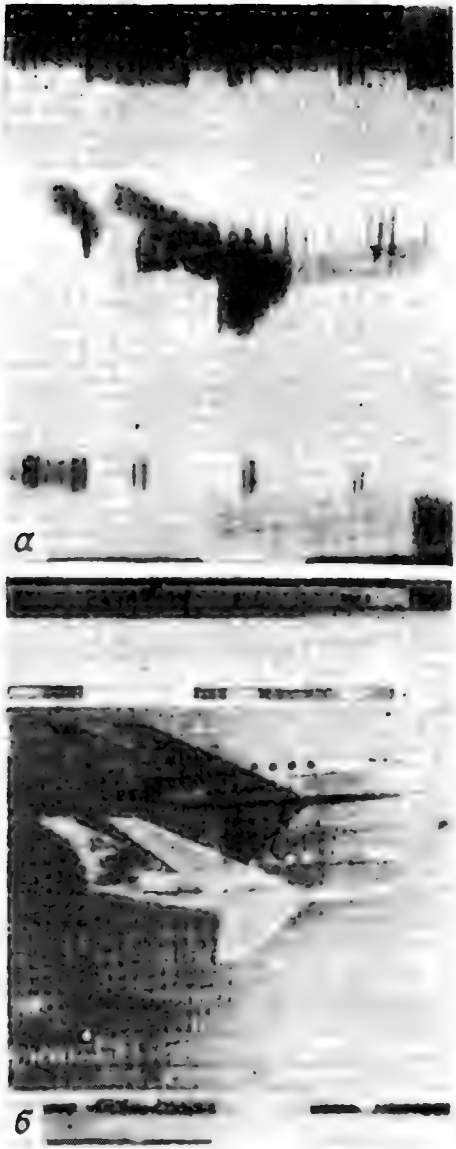
Следовательно, выходное изображение $\hat{o}(x)$ содержит два смещенных варианта объекта повышенной резкости, которые соответствуют функциям рассеяния, представленным положительным и отрицательным пиками на фиг. 5.4. Пример применения этого метода иллюстрирует фиг. 5.5. Хотя в данном случае была использована аналоговая технология, метод вполне подходит для цифровой обработки.

Искажения, вызываемые движением, могут быть устранены также путем пространственного дифференцирования изображения $i(x)$. Однако в специальном исследовании Хомена [33] было показано, что применение маски Суинделла приводит к меньшему увеличению шума на выходе, чем метод дифференцирования.

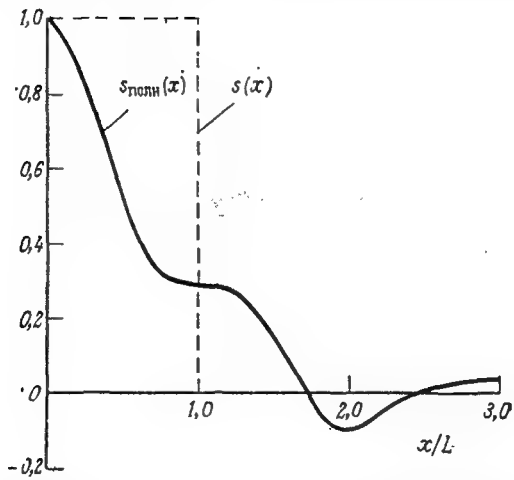
5.10.3. Чистая фазовая компенсация

Спектральные составляющие объекта $O(\omega)$ с частотами, на которых функция передачи системы отображения $\tau(\omega)$ отрицательна, смещаются по изображению в поперечном направлении.

[1]) Здесь исправлена неточность в оригинале. — *Прим. перев.*



Фиг. 5.5. Применение метода, принцип которого объяснен на фиг. 5.4 *а* — изображение сильно искажено в результате прямолинейного движения в вертикальном направлении. *б* — два улучшенных варианта изображения самолета, полученные путем противосвертки с маской $m(x)$. Функция маски $m(x)$ была сделана достаточно протяженной, чтобы обеспечить разделение выходных изображений. Как видно из фотографий, значительная часть деталей, размытых в исходном изображении *а*, восстановлена в *б* [32].



Фиг. 5.6. Функция рассеяния $s_{полн}(x)$, полученная путем чистой компенсации фазы спектра, соответствующего прямолинейному движению объекта на расстояние $2L$. Компенсация обрывается на втором нуле функции передачи $\mathrm{sinc}(\omega L)$, соответствующей движению изображения. В результате компенсации область между первым и вторым нулем просто «поворачивается» из отрицательного направления в положительное. Как видно из сравнения $s_{полн}(x)$ с исходной функцией $s(x)$, соответствующей смазыванию, достигаемый выигрыш в разрешении невелик. Следовательно, как метод реставрации компенсация фазы малоэффективна.

Эти составляющие взаимодействуют с несмещенными составляющими, частоты которых соответствуют положительным зонам $\tau(\omega)$. В результате возникает впечатление смазанного изображения.

Было сделано предположение, что для процессов отображения, у которых $\tau(\omega)$ имеет значительные отрицательные области, такой фазовый тип смазывания является основной причиной искажения изображения и действует значительно сильнее, чем чистое ослабление $\tau(\omega)$ на высоких частотах. Представлялось, такое искаженное изображение можно улучшить просто путем внесения в $I(\omega)$ опережения или отставания по фазе, которое скомпенсирует фазовые искажения, вызываемые $\tau(\omega)$. Таким образом, если $\tau(\omega) \equiv |\tau(\omega)| \exp[j\phi(\omega)]$, то улучшение можно обеспечить формированием

$$\hat{O}(\omega) = I(\omega) \exp[-j\phi(\omega)]. \qquad (5.57)$$

При такой обработке получаем $\tau_{полн}(\omega) = |\tau(\omega)|$ и, следовательно,

$$s_{полн}(x) = (2\pi)^{-1/2} \int_{-\Omega_p}^{\Omega_p} d\omega\, |\tau(\omega)| \exp(+j\omega x). \qquad (5.58)$$

Поскольку для прямолинейного движения изображения $\tau(\omega)$ получается как ПФ от $\mathrm{Rect}(x/L)$ в виде $\tau(\omega)=2L\,\mathrm{sinc}(\omega L)$ и поскольку функция $\mathrm{sinc}(\omega L)$ имеет значительные отрицательные области, обработка изображения описанным способом должна была дать эффективную компенсацию искажений, вызываемых движением. Однако оказалось, что это не вполне верно. При экспериментах на ЭВМ [34] выяснилось, что улучшение изображения, достигаемое с помощью алгоритма (5.57), ненамного сильнее, чем при использовании чистой инверсной фильтрации.

Чтобы подтвердить этот результат, мы показали на фиг. 5.6 полную функцию рассеяния, вычисленную для этого эксперимента при условии, что граничная частота обработки Ω_p соответствует второму нулю функции $\mathrm{sinc}(\omega L)$. При этом операция (5.57) приводит к переворачиванию в положительную сторону первого (и наибольшего) отрицательного лепестка функции $\mathrm{sinc}(\omega L)$. Из фиг. 5.6 видно, что $s_{\text{полн}}(x)$ по существу состоит из суммы двух ФРТ: одной с первым нулем в $x_0 \approx L$ и другой с первым нулем в $x_0 \approx 2L$. Мы заключаем, что чистую фазовую компенсацию следует производить лишь тогда, когда требуется небольшое улучшение (с соответствующим небольшим увеличением шума).

5.10.4. Метод последовательной свертки ван Циттерта

Мы убедились с помощью (5.31), что линейная фильтрация может быть полностью выполнена в исходной пространственной области переменной x. Функцию $y(x)$ в (5.31) часто называют «противосверточной» (deconvolving). Очевидно, если $y(x)$ везде положительна, то выход $\hat{o}(x)$ неизбежно будет даже более сглаженным, т. е. более искаженным, чем входное изображение. Следовательно, для улучшения изображения функция $y(x)$ должна содержать значительные отрицательные области.

Предположим, что пользователь хочет выполнить улучшение с помощью линейной фильтрации в области переменной x, но должен ограничиться использованием чисто положительной функции $y(x)$; это ограничение может быть обусловлено, например, используемым аналоговым прибором. В соответствии с предыдущим рассуждением это представляется невозможным. Однако приведенные соображения относятся к использованию однократной свертки $i \otimes y$. Если же пользователь может производить многократную свертку, то появляется возможность использовать положительные ядра $y(x)$. Интересно, что при этом $s(x)$ выполняет функцию $y(x)$, как установил в своем исследовании ван Циттерт [35] в 1931 г. Джанссон [10] рассмотрел некоторые современные спектроскопические приложения метода ван Циттерта. Этот метод состоит в следующем.

Пусть индекс (k) указывает номер итерации. Начнем с $k=0$ и исходного $\hat{o}^{(0)}(x)=i(x)$ и применим следующий алгоритм:

$$\begin{aligned}
&1)\ \hat{i}^{(k)}(x)=\hat{o}^{(k)}(x)\otimes s(x),\\
&2)\ \hat{o}^{(k+1)}(x)=\hat{o}^{(k)}(x)+[i(x)-\hat{i}^{(k)}(x)],\\
&3)\ k\to k+1,\\
&4)\ \text{перейти к (1).}
\end{aligned}\tag{5.59}$$

Опытным путем было обнаружено, что метод ван Циттерта не всегда сходится и особенно в тех случаях, когда изображение $i(x)$ сильно смазано или когда в $i(x)$ имеется значительный шум [36]. Далее мы объясним, почему это происходит, и покажем, что метод ван Циттерта при $k\to\infty$ фактически сходится в пределе *к оценке* (5.35), *получаемой с помощью инверсной фильтрации*, при условии, что $|1-\tau(\omega)|<1$ для всех ω. Кроме того, мы покажем, что на частотах ω, для которых $\tau(\omega)=0$, шум в изображении при последовательных итерациях *линейно нарастает*!

Анализ оказывается очень простым, если перевести (5.59) в частотную область. В результате после некоторого комбинирования получаем (опуская аргумент ω)

$$\hat{O}^{(k+1)}=\hat{O}^{(k)}\cdot(1-\tau)+I.$$

Начав с $\hat{O}^{(0)}=I$ в правой части и подставляя последовательно левую часть в правую, получаем геометрическую прогрессию

$$\hat{O}^{(k)}=I\sum_{j=0}^{k}(1-\tau)^j,$$

которая вычисляется аналитически и составляет

$$\hat{O}^{(k)}=(I/\tau)[1-(1-\tau)^{k+1}].\tag{5.60}$$

Этот результат, верный для любой частоты ω, показывает, что:

1. В пределе при $k\to\infty$ $\hat{O}^{(k)}(\omega)$ приближается к оценке с помощью инверсного фильтра $I(\omega)/\tau(\omega)$ при условии, что $|1-\tau(\omega)|<1$. Следовательно, если, например, $\tau(\omega)<0$ вследствие прямолинейного движения изображения, дефокусировки или смазывания по другим причинам, то $\hat{O}^{(k)}(\omega)$ не сходится и поэтому не будет преобразованием Фурье от оценки $\hat{o}^{(k)}(x)$, которая входит в алгоритм (5.59) как наблюдаемая величина.

2. При конечном k функция в квадратных скобках в (5.60) действует как аподизирующая, или сглаживающая, функция окна, подобная функциям, описанным в п. 5.9.2. Следовательно, при конечных k метод ван Циттерта приводит к сглаженному варианту инверсной фильтрации. С ростом k функция окна приближается к $\mathrm{Rect}(\omega/\Omega)$.

3. Переходя к пределу при $\tau \to 0$ и используя правило Лопиталя, находим, что на частотах, на которых $\tau(\omega) = 0$, оценка выражается в виде $\hat{O}^{(k)}(\omega) = (k+1)N(\omega)$, что соответствует просто линейно усиленному шуму.

4. Наконец, при малых τ и средних k в (5.60) мы обнаруживаем, что оценка $\hat{O}^{(k)}(\omega)$ соответствует просто линейно усиленному варианту $(k+1)I(\omega)$ входного изображения.

По-видимому, эти результаты объясняют описанные выше наблюдения Джанссона [36], указывающие на узкие рамки применения метода ван Циттерта. Особое значение имеет требование никогда не допускать $\tau(\omega) < 0$ и избегать явления (4); последнее предупреждение весьма важно, поскольку для всех оптических изображений $\tau(\omega)$ становится малым вблизи граничной частоты Ω.

Учитывая недостатки этого метода, поразительно, что Джанссону и др. [10] удалось найти простой прием, позволяющий добиваться успеха при обработке реальных сильно смазанных изображений (раздел 5.16).

5.10.5. Методы дискретной противосвертки

Несмотря на высокую скорость алгоритма БПФ (раздел 5.7), некоторые исследователи считают полезным выполнять цифровую обработку в пространственной области, поскольку в принципе операция противосвертки (5.31) может быть выполнена с помощью дискретного суммирования, охватывающего несколько точек. Это можно показать следующим образом.

Для того чтобы использовать выражение (5.31) в ЦВМ, его нужно преобразовать в дискретную форму:

$$\hat{o}(x_n) = \sum_{m=1}^{M} y_m i(x_n - x_m), \quad n = 1, 2, \ldots, N, \tag{5.61}$$

где точки x_m входных данных разнесены, скажем, на интервал Найквиста (раздел 5.2). Будем теперь рассматривать функцию ядра противосвертки y как вектор из M входных чисел, выбираемых, как будет показано далее.

Изображение i связано с объектом o посредством (5.1). Подставляя (5.1) в (5.61), получаем

$$\hat{o}(x_n) = \int_{-X}^{X} dx\, o(x)\, s_{\text{полн}}(x - x_n), \tag{5.62а}$$

где полная (от входа до выхода системы) ФРТ имеет вид

$$s_{\text{полн}}(x) \equiv \sum_{m=1}^{M} y_m s(x - x_m). \tag{5.62б}$$

Как следует из (5.62а), желательно, чтобы $s_{\text{полн}}$ как можно лучше аппроксимировала дельта-функцию Дирака $\delta(x)$. С другой стороны, согласно (5.62б), $s_{\text{полн}}$ выражается взвешенной суммой отсчетных значений известной ФРТ. Поэтому веса y_m должны выбираться так, чтобы взвешенная сумма обеспечивала наилучшее (в некотором смысле) приближение к дельта-функции.

Для примера предположим, что имеется дифракционно-ограниченная система отображения, у которой

$$s(x) = \operatorname{sinc}^2(x). \tag{5.63а}$$

При этом полоса частот Ω, интервал Найквиста и положение первого нуля x_0 составляют соответственно 2, $\pi/2$ и π. Установим смещения изображения x_m в (5.61) кратными интервалу Найквиста.

Потребуем, чтобы в результате расчета $s_{\text{полн}}(x)$ в (5.62б) имела первый нуль при $x_0 = \pi/2$, т. е. вдвое ближе к началу, чем у физической ФРТ (5.63а), и будем считать, что этого результата мы хотим добиться, используя лишь $M = 3$ члена в (5.61) и (5.62б). С учетом этих ограничений и того, что $s_{\text{полн}}(x)$ должна быть симметричной относительно x, (5.62б) принимает вид

$$s_{\text{полн}}(x) = \operatorname{sinc}^2(x) + y_1[\operatorname{sinc}^2(x - \pi/2) + \operatorname{sinc}^2(x + \pi/2)]. \tag{5.63б}$$

Вес y_1 должен удовлетворять условию разрешения $s_{\text{полн}}(\pi/2) = 0$, откуда следует

$$0 = \operatorname{sinc}^2(\pi/2) + y_1[\operatorname{sinc}^2(0) + \operatorname{sinc}^2(\pi)]. \tag{5.63в}$$

Это уравнение имеет единственное решение

$$y_1 = -\operatorname{sinc}^2(\pi/2) = -4/\pi^2. \tag{5.64}$$

В результате получаем, что для данного случая ($M = 3$) сумма изображений со смещениями $\{x_m\} = (0, -\pi/2, +\pi/2)$ и весами $(1, -4/\pi^2, +4/\pi^2)$ образует полную ФРТ, первый нуль которой занимает положение на оси x вдвое ближе, чем первый нуль физической ФРТ.

Чтобы оценить этот результат, нужно сравнить его по точкам с $s_{\text{полн}}$, получаемой при инверсной фильтрации, которая, согласно (5.38), составляет в данном случае $\operatorname{sinc}(2x)$. Эта функция также имеет первый нуль при $x_0 = \pi/2$, т. е. в месте, соответствующем нашему решению (5.63б) и (5.64). Однако график (5.63б) имеет боковой лепесток величиной 0,39, что превышает величину бокового лепестка $\operatorname{sinc}(2x)$, составляющую 0,21. Следовательно, инверсная фильтрация обеспечивает лучший результат, чем дискретная противосвертка при $M = 3$.

Это положение исправляется при увеличении M. В общем случае способ, принятый при составлении (5.63б), приводит к

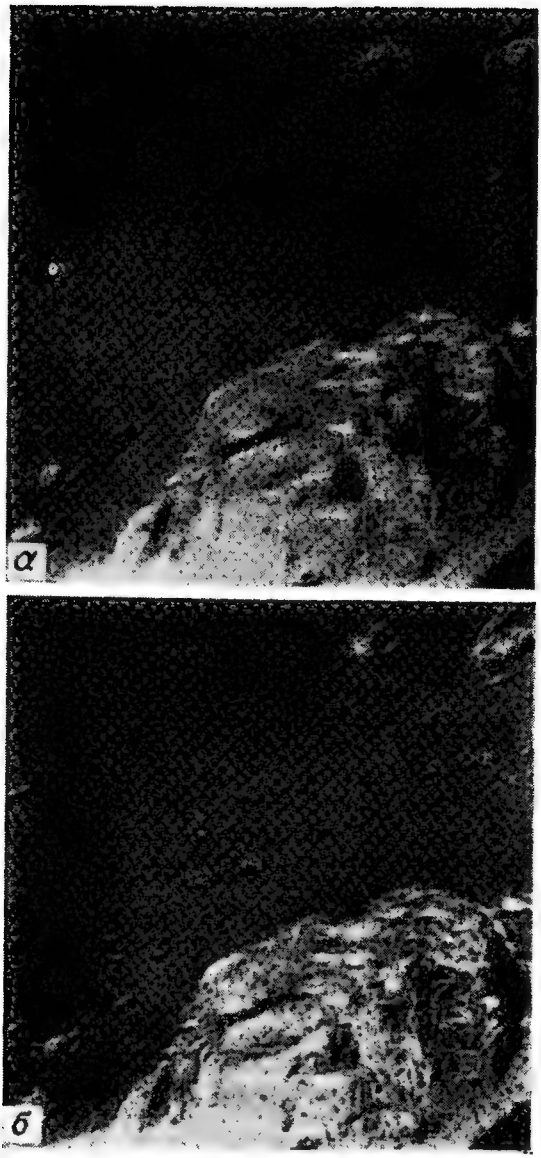
появлению $(M-1)/2$ независимых $\{y_m\}$. Поскольку имеется лишь одно ограничительное уравнение $s_{\text{полн}}(x_0)=0$, которому должны удовлетворять $\{y_m\}$, то $(M-3)/2$ из них оказываются произвольными параметрами. Их можно подобрать так, чтобы в каком-то смысле минимизировать боковые лепестки $s_{\text{полн}}(x)$ (см. [37]). Например, если требуемое разрешение по-прежнему составляет $x_0=\pi/2$, то при $M=7$ боковые лепестки получаются почти такими же, как при инверсной фильтрации. С другой стороны, если для пользователя приемлемы величины боковых лепестков, превышающие 0,21, он может получить более высокое разрешение, чем достигается при инверсной фильтрации (см. [37]).

Наконец, в [37] было установлено, что первый нуль при $x_0=\pi/2$ может быть получен при использовании только трех целых значений весов $\{y_m\}=(0, -1, +1)$. В соответствии с (5.61) это приводит к построению способа реставрации, включающего только чистое сложение и вычитание соседних значений изображения.

Дискретный подход к обработке данных, соответствующий (5.61), был принят также Аргуелло и др. [28]. Однако эти авторы при выборе коэффициентов $\{y_m\}$ исходили из функций окна $W(\omega)$, определяемых в частотной области. Такой подход принципиально и количественно отличается от описанного выше метода проектирования в исходной (пространственной) области в соответствии с (5.63) и (5.64). Здесь уместно спросить, что оказывает большее воздействие на качество изображения — контраст или рассеяние точки? Вот один пример: наблюдая поведение кривой $W(\omega)$ общего вида, нельзя судить о величине наибольшего бокового лепестка и положении первого нуля ее преобразования Фурье $s_{\text{полн}}(x)$.

Тем не менее, как и в работе [37], в работе Аргуелло и др. делается вывод, что для получения приемлемой реставрации достаточно брать число слагаемых M порядка трех — пяти. Следовательно, исходя из двух различных точек зрения, можно заключить, что реставрировать данные с использованием обработки ЭВМ в исходной области пространственных переменных практически также полезно, как и с использованием обработки в частотной области.

Такие наиболее широко известные примеры улучшения изображения, как обработки снимков поверхности Луны, полученных с автоматической межпланетной станции «Рейнджер», и других фотографий, выполненные Натаном [38] и его сотрудниками, фактически производились с использованием противосвертки (5.61) в области пространственных координат. На фиг. 5.7 показан характерный пример фотографий до и после обработки по методу Натана.



Фиг. 5.7. *а* — изображение поверхности Луны, полученное с автоматической межпланетной станции «Рейнджер VIII», до компенсации функции передачи. *б* — изображение после компенсации. Функция окна $W(\omega)$ для этой реставрации была равна единице (что соответствует полной компенсации) до частоты, где $\tau(\omega)$ снижалась до 0,2; от этой точки $W(\omega)=5\tau(\omega)$, так что далее $W(\omega)$ плавно приближалась к нулю. Операции по реставрации на ЦВМ производились в области пространственной переменной в соответствии с (5.61) [38].

5.11. Линейные методы, основанные на статистическом подходе

Произвольный характер функции окна $W(\omega)$, о котором говорилось выше, побудил ряд авторов попытаться построить производную функцию окна на основе какого-либо перспективного критерия, каким является, например, минимум среднеквадратичной ошибки (МСКО). В таком случае цель применения функции $W(\omega)$ состоит в снижении паразитной осцилляции в $\hat{o}(x)$, вызываемой «звенящей» реакцией системы и наличием случайного шума в данных. Последнее обстоятельство наводит на мысль воспользоваться вероятными (ожидаемыми) значениями.

5.11.1. Метод фильтрации по Винеру с учетом ограничения резкости

Указанные факторы побудили Хелстрома [39] использовать подход, основанный на применении фильтрации по Винеру. При этом функция передачи реставрирующего фильтра $Y(\omega)$, на которую умножается преобразование Фурье данных $I(\omega)$, должна удовлетворять критерию качества МСКО в виде

$$\left\langle \int_{-\Omega}^{\Omega} d\omega \, | O(\omega) - \hat{O}(\omega) |^2 \right\rangle = \min, \tag{5.65а}$$

где

$$\hat{O}(\omega) \equiv \begin{cases} Y(\omega) \cdot I(\omega) & \text{при } |\omega| \leqslant \Omega, \\ 0 & \text{при } |\omega| > \Omega. \end{cases} \tag{5.65б}$$

(Бакус и Гильберт [42] применили несколько иной подход, который будет описан ниже.)

Задача отыскания $Y(\omega)$ в представленной здесь постановке легко решается, как показано в работе [39]. Однако интересно выяснить, что произойдет, если при реставрации мы присоединим к критерию (5.65а) ограничение общей резкости S, определяемой как

$$S \equiv \left\langle \int_{-\infty}^{\infty} dx \, | \, d\hat{o}(x)/dx \, |^2 \right\rangle = \tag{5.66а}$$

$$= \left\langle \int_{-\Omega}^{\Omega} d\omega \, \omega^2 \, | \, \hat{O}(\omega) \, |^2 \right\rangle. \tag{5.66б}$$

Второе выражение получено с использованием тождеств (5.5), (5.10) и (5.12). Резкость здесь служит мерой общего содержания контурных градиентов в выходе $\hat{o}(x)$.

Таким образом, новый, более полный критерий можно выразить в виде

$$\int_{-\Omega}^{\Omega} d\omega \langle | O(\omega) - Y(\omega) I(\omega) |^2 \rangle + \lambda \int_{-\Omega}^{\Omega} d\omega \omega^2 \langle | Y(\omega) I(\omega) |^2 \rangle = \min. \quad (5.67)$$

Из (5.67) видно, что степень воздействия резкости на ожидаемую ошибку (и на согласованность с входными данными) определяется параметром λ. Этот параметр выполняет функцию, в известной мере аналогичную функции коэффициента сглаживания γ в алгоритме Филлипса (5.15).

Если в (5.67) раскрыть члены с квадратами модулей, используя уравнение отображения (5.29) и предположение о статистической независимости объекта и шума

$$\langle N^*(\omega) O(\omega) \rangle = 0,$$

мы получим

$$\int_{-\Omega}^{\Omega} d\omega [\phi_0 - Y^*\tau^*\phi_0 - Y\tau\phi_0 + YY^*(|\tau|^2\phi_0 + \phi_n) + \lambda\omega^2 YY^*(|\tau|^2\phi_0 + \phi_n)] = \min,$$

где ϕ_0 и ϕ_n — энергетические спектры объекта и шума, определяемые согласно (5.4а). Ради краткости частотные аргументы всех функций в квадратных скобках опущены.

Теперь нетрудно видеть, что данная задача представляет собой задачу типа Эйлера — Лагранжа, причем выражение в квадратных скобках есть лагранжиан $\mathscr{L}$, который не является функцией $\partial Y/\partial\omega$ или $\partial Y^*/\partial\omega$. Поэтому решение выражается просто в виде

$$\partial\mathscr{L}/\partial Y^* = \partial\mathscr{L}/\partial Y = 0.$$

Обе производные дают одинаковый результат

$$Y(\omega) = \left(\frac{\tau^*\phi_0}{|\tau|^2\phi_0 + \phi_n}\right)\left(\frac{1}{1+\lambda\omega^2}\right). \quad (5.68)$$

Обсуждение

Из полученной формулы видно, что фильтр, построенный на основе комбинированного критерия МСКО — резкость, естественным образом распадается на два последовательных фильтра, которые действуют именно в этом порядке. Первый фильтр является реставрирующим по Хелстрому [39]; второй может рассматриваться как фильтр управления резкостью, поскольку при $\lambda > 0$ он приводит к ослаблению высоких частот и тем самым к сглаживанию выхода $\hat{o}(x)$, тогда как при $\lambda < 0$ он вызы-

вает подъем высоких частот. Эксперименты, проведенные с различными λ, показали, что полезные результаты получаются при $\lambda > 0$, т. е. в режиме сглаживания [34].

Поскольку $O_{\text{снг}} = YI_{\text{снг}} = (Y\tau)O$, полученный фильтр (5.68) соответствует производной функции окна $W = \tau \cdot Y$. Именно такую цель мы ставили вначале.

Первая дробь в скобках в (5.68) — реставрирующий фильтр Хелстрома. Когда $\phi_n/\phi_0 \to 0$, этот фильтр переходит в инверсный, т. е. дробь принимает вид $1/\tau(\omega)$; наоборот, когда $\phi_n/\phi_0 \to \infty$, дробь приближается к нулю (происходит подавление шума). Эти свойства, очевидно, соответствуют ожидаемым характеристикам фильтра, минимизирующего выходную ошибку в присутствии шума.

Хотя фильтр (5.68) обладает высокой степенью гибкости и «встроенной» способностью подавления шума, у него имеются недостатки, перечисленные в разделе 5.9. Кроме того, энергетические спектры $\phi_0(\omega)$ и $\phi_n(\omega)$ практически редко известны, а использованный МСКО-критерий (5.65а) имеет свои собственные недостатки. Следует заметить, что критерий (5.65а) обеспечивает малую ошибку лишь в среднем, т. е. по множеству реставраций, соответствующих известным энергетическим спектрам (см. выше); пользователю же требуется уверенность в качестве всего одной или двух конкретных реставраций. Сама исходная посылка, что с помощью фильтра (5.68) должно быть выполнено большое количество реставраций, часто оказывается неверной в случае улучшения изображений.

Информационный аспект оптимальной фильтрации

Рассмотрим теперь уравнение (5.68) в случае $\lambda = 0$, когда, согласно (5.67), обеспечивается абсолютный минимум СКО. Довольно малоизвестно, что $Y(\omega)$ в (5.68) прямо связано с количеством информации (по Шеннону), содержащейся в спектральных данных об изображении $I(\omega)$. Фельжет и Линфут в своей классической статье [40] установили, что максимум информации по Шеннону $SI(\omega)$ на данной частоте составляет

$$SI(\omega) = \log\left[1 + \frac{|\tau(\omega)|^2 \phi_0(\omega)}{\phi_n(\omega)}\right]. \tag{5.69}$$

Эту величину называют также пропускной способностью канала для процесса формирования изображения. Комбинируя (5.68) и (5.69), мы получаем выразительный результат:

$$Y(\omega) = [1/\tau(\omega)] \cdot [1 - e^{-SI(\omega)}]. \tag{5.70}$$

Это соотношение непосредственно показывает, как информация ограничивает способность фильтра производить оптимальную

реставрацию линейным способом. Здесь коэффициент полной компенсации $1/\tau$ просто модулируется функцией от информации SI. Мы видим, что полная компенсация обеспечивается оптимальным фильтром только при бесконечной информации, и наоборот, при отсутствии информации компенсация невозможна.

Покажем, что МСКО также ограничивается количеством информации. При $\lambda = 0$ использование оптимального фильтра $Y(\omega)$ в (5.67) дает минимизированную СКО

$$\varepsilon^2 = \int_{-\Omega}^{\Omega} d\omega \frac{\phi_n(\omega)\,\phi_0(\omega)}{|\tau(\omega)|^2\,\phi_0(\omega) + \phi_n(\omega)}. \tag{5.71}$$

Используя снова выражение (5.69), получаем

$$\varepsilon^2 = \int_{-\Omega}^{\Omega} d\omega\,\phi_0(\omega)\,e^{-SI(\omega)}. \tag{5.72}$$

Поскольку $\phi_0(\omega) \geqslant 0$, отсюда следует, что МСКО убывает экспоненциально с увеличением количества информации (см. [41]).

5.11.2. Метод оптимизации Бакуса — Гильберта

При $\lambda = 0$ критерий (5.67) позволяет выполнять однопараметрическую оптимизацию, причем параметром служит МСКО. Однако это не единственный полезный параметр. Укажем для примера один возможный недостаток критерия (5.67): получаемый фильтр Y может слишком сильно ослабить входы $I(\omega)$ из-за относительно больших значений входной функции $\phi_n(\omega)$. Вследствие этого выход $\hat{o}(x)$ может оказаться смазанным настолько сильно, что не удовлетворит требованиям пользователя. Хотя эту трудность можно преодолеть, делая произвольный параметр λ отрицательным (как было указано выше), однако возможен и другой путь.

Бакус и Гильберт [42] предложили минимизировать (в некотором рациональном смысле) *в отдельности*: 1) выходной шум и 2) выходное отклонение реставрации сигнала от истинного объекта. В качестве меры первой величины пользователь может применить, например, среднеквадратичное значение полного выходного шума, которое, согласно (5.30б), составляет

$$\eta^2 \equiv \left\langle \int_{-\infty}^{\infty} d\omega\,|NY|^2 \right\rangle. \tag{5.73а}$$

Обычной мерой второй величины является среднеквадратичное отклонение выхода фильтра $Y\tau O$ в соответствии с (5.30б) от

требуемой оценки (с учетом функции окна) $W \cdot O$:

$$\mu^2 \equiv \left\langle \int_{-\infty}^{\infty} d\omega \, | W(\omega) O(\omega) - Y(\omega) \tau(\omega) O(\omega) |^2 \right\rangle. \qquad (5.73б)$$

Немного ниже мы воспользуемся этими специальными мерами.

Другая возможная мера второй величины, которую фактически использовали Бакус и Гильберт, имеет вид

$$l^2 \equiv \int_{-\infty}^{\infty} dx \, x^2 [s_{\text{полн}}(x)]^2. \qquad (5.74)$$

Заметим, что чем меньше (5.74), тем ближе $s_{\text{полн}}(x)$ к дельта-функции $\delta(x)$. В работе [42] показано, что уравнение комбинированного критерия

$$\lambda_1 \eta^2 + \lambda_2 l^2 = \min \qquad (5.75)$$

может быть решено относительно противосверточной функции $y(x)$ при различных сочетаниях входов λ_1 и λ_2. Эти решения выражаются в виде кривых взаимосвязи между достижимым (наикратчайшим) расстоянием разрешения l, равным квадратному корню из (5.74), и допустимым уровнем шума η, который должен быть ниже требуемого фиксированного уровня.

Однако этот метод имеет следующие недостатки:

1. В силу линейности выход $\hat{o}(x)$ является неограниченным и поэтому может иметь отрицательные области.

2. Мера (5.74), служащая для улучшения разрешения, довольно произвольна и в значительной степени нечувствительна к структуре боковых лепестков в $s_{\text{полн}}(x)$. Поэтому выражение для достижимого расстояния разрешения l не содержит информации о нежелательной осцилляции, которую можно ожидать на выходе.

3. Для изображения, содержащего $M \times M$ входных значений, этот метод требует решения системы линейных уравнений размером $M^2 \times M^2$. Следовательно, даже при изображении среднего размера, скажем 32×32, возникают существенные вычислительные трудности. До настоящего времени не появилось сообщений об испытании метода в приложении к двумерному отображению; для одномерного отображения он был испытан Салехом [43].

Вычислительные трудности, очевидно, можно преодолеть, воспользовавшись методом фильтрации. Для этого внесем видоизменение в критерий (5.75), заменив меру (5.74) мерой (5.73б). В соответствии с этим будем искать фильтр Бакуса — Гильберта $Y_{\text{БГ}}(\omega)$, удовлетворяющий уравнению

$$\lambda_1 \eta^2 + \lambda_2 \mu^2 = \min, \qquad (5.76)$$

Параметры λ_1 и λ_2 выбираются пользователем так, чтобы либо усилить подавление шума (большое λ_1/λ_2), либо улучшить разрешение (малое λ_1/λ_2).

Подставляя (5.73а) и (5.73б) в (5.76), возводя модули в квадрат и беря затем производную $\partial/\partial Y^*$, как при выводе (5.68), находим решение в виде

$$Y_{\text{BG}}(\omega) = \frac{\tau^*(\omega)\,\phi_0(\omega)\,W(\omega)}{|\tau(\omega)|^2\,\phi_0(\omega) + (\lambda_1/\lambda_2)\,\phi_n(\omega)}. \tag{5.77}$$

Это решение имеет явное сходство с винеровским реставрирующим фильтром (5.68) при снятом ограничении $\lambda = 0$. С другой стороны, (5.77) обладает требуемыми свойствами, обеспечивая при малых λ_1/λ_2 увеличение разрешения (причем $Y_{\text{BG}} \to W/\tau$, т. е. в пределе получается инверсный фильтр, взвешенный функцией окна), а при больших λ_1/λ_2 — усиление подавления шума.

Таким образом, принцип Бакуса — Гильберта, состоящий в раздельном взвешивании разрешения и шума, можно рассматривать как способ фильтрации, легко приспосабливаемый для решения задач улучшения двумерных изображений. В сочетании с алгоритмом БПФ фильтр (5.77) обеспечивает гибкий способ реставрации изображений большого размера.

Однако этот способ, как и стандартный способ Бакуса — Гильберта, описанный в работе [42], имеет недостатки — он допускает появление нефизических (отрицательных) областей на выходе, и в нем отсутствуют средства прямого управления боковыми лепестками $s_{\text{полн}}(x)$, которая представляет собой ПФ от τY_{BG}.

5.12. Возможности экстраполяции полосы частот

Стандартный метод фильтрации, определяемый выражениями (5.30а) и (5.30б), ограничивает спектральные операции пределами конечной полосы частот $|\omega| \leqslant \Omega_p$, причем $\Omega_p < \Omega$ (Ω — оптическая граничная частота). С другой стороны, опытный пользователь может заметить, что иногда реставрированный спектр $\hat{O}(\omega)$ оказывается весьма гладкой функцией в пределах обрабатываемой полосы частот $|\omega| \leqslant \Omega_p$. Если принять не лишенное смысла допущение, что $\hat{O}(\omega)$ остается гладким на частотах $|\omega| > \Omega_p$, то у пользователя может появиться стимул экстраполировать $\hat{O}(\omega)$ за пределы абсолютной граничной частоты Ω, так что условие $0 \leqslant |\omega| \leqslant \Omega_p$ будет выполняться для $\Omega_p > \Omega$. Если бы экстраполяцию удалось выполнить корректно, то это позволило бы получить выигрыш в разрешении, пропорциональный расширению полосы частот; см. (5.40). Насколько корректна подобная экстраполяция? Насколько сильная экстраполяция требуется для достижения измеримого эффекта?

Чтобы подойти к решению этих проблем, предположим для простоты, что объект состоит из двух точек одинаковой яркости с расстоянием a между ними:

$$o(x) = \delta(x - a/2) + \delta(x + a/2). \tag{5.78а}$$

Обозначим через $\hat{o}_{\text{экст}}(x)$ реставрированный выход, полученный с помощью идеальной спектральной реставрации $\hat{O}(\omega) = O(\omega)$, $0 \leqslant |\omega| \leqslant \Omega_p$ при $\Omega_p > \Omega$. Далее рассмотрим, насколько сильная экстраполяция, измеряемая отношением Ω_p/Ω, необходима для того, чтобы разрешить две точки на выходе $\hat{o}_{\text{экст}}(x)$. Полагая, что наблюдение $\hat{o}_{\text{экст}}(x)$ ведется невооруженным глазом, будем исходить из критерия Релея и считать две точки разрешенными, когда глубина седловины между ними составляет приблизительно 20%.

Расстояние a полезно представить как долю r релеевского расстояния разрешения $2\pi/\Omega$:

$$a \equiv r(2\pi/\Omega). \tag{5.78б}$$

Будем считать, что чем меньше r, тем больше требуемое относительное расширение полосы частот Ω_p/Ω для разрешения двух точек.

Спектр объекта находится как ПФ от (5.78а) и имеет вид

$$O(\omega) = (2\pi)^{1/2} \cos(\pi r\omega/\Omega). \tag{5.78в}$$

Предполагается, что этот спектр полностью известен во всей обрабатываемой полосе частот $0 \leqslant |\omega| \leqslant \Omega_p$. Поэтому для получения итоговой оценки подставим (5.78в) в формулу обратного преобразования (5.30в), откуда получим

$$\hat{o}_{\text{экст}}(x) = (\pi)^{-1}\,\Omega_p\{\operatorname{sinc}[\Omega_p(x + \pi r/\Omega)] + \operatorname{sinc}[\Omega_p(x - \pi r/\Omega)]\}. \tag{5.78г}$$

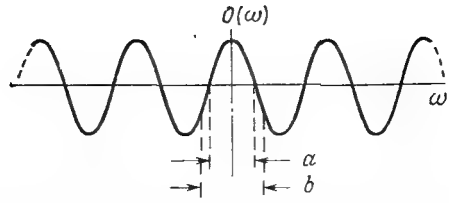
В соответствии со сказанным выше будем считать, что для разрешения двух точек $\hat{o}_{\text{экст}}(x)$ должна иметь при $x = 0$ седловину глубиной по меньшей мере 20% по отношению к значениям при $x = \pm a/2$, или, учитывая тождество (5.78б), при $x = \pm r\pi/\Omega$. Согласно (5.78г), искомое отношение составляет

$$\hat{o}_{\text{экст}}(0)/\hat{o}_{\text{экст}}(a/2) = \frac{2\operatorname{sinc}(\pi r\Omega_p/\Omega)}{1 + \operatorname{sinc}(2\pi r\Omega_p/\Omega)}. \tag{5.78д}$$

Путем непосредственной подстановки в (5.78д) пробных значений $r\Omega_p/\Omega$ определяем, что требуемая седловина глубиной в 20% обеспечивается приблизительно при

$$\Omega_p/\Omega = 3/(4r).$$

Отсюда получаем интересные примеры: если $r = 1/2$ (т. е. точки объекта разнесены на половину релеевского расстояния разрешения), то требуемое относительное расширение полосы частот



Фиг. 5.8. Спектр двухточечного объекта в виде косинусоиды. Какой участок косинусоиды требуется для разрешения двух точек, если используется преобразование Фурье этого участка? Фактически достаточно участка шириной b. Участок шириной a соответствует полосе частот оптики. Его преобразование Фурье не обеспечивает разрешения двух точек. Однако для того, чтобы добиться разрешения этих точек, требуется сгладить спектральную кривую и увеличить ширину рабочего участка всего лишь от a до b.

$\Omega_p/\Omega = {}^3/_2$ (необходимо расширение на 50%); если же $r = {}^1/_3$, то требуемое расширение составляет ${}^9/_4$, т. е. значительно выше, чем в первом примере.

Все требуемые экстраполяции (для различных r) фактически имеют границей одну и ту же точку косинусоиды (5.78в), выражающей спектр объекта. Подставляя Ω_p из (5.79) вместо ω в (5.78в), находим значение

$$O(\omega) = \cos(3\pi/4), \tag{5.80}$$

которое соответствует точке несколько дальше первого нуля косинусоиды. Следовательно, чтобы разрешить точки двухточечного объекта, в оценку $\hat{o}_{\text{экст}}(x)$ должна быть введена лишь весьма небольшая доля всего спектра ($|\omega| \leqslant \infty$), описывающего такой объект. Кроме того, этот спектр (косинусоида) весьма гладок (почти линеен) до значения (5.80). Описанный случай иллюстрирует фиг. 5.8.

Таким образом, при решении задачи разрешения двухточечного объекта требуемая экстраполяция простирается лишь несколько дальше первого нуля спектра объекта (выражаемого косинусоидой); экстраполяция захватывает только весьма гладкую, почти линейную часть спектра; чтобы обеспечить разрешение объектов на расстоянии, составляющем около ${}^1/_2$ релеевского расстояния разрешения, требуется расширение полосы частот приблизительно лишь на 50%.

Было бы неправильно объявить, что изложенное доказывает реализуемость методов экстраполяции. Однако мы показали, что метод экстраполяции, позволяющий получить гладкое продолжение данной кривой $\hat{O}(\omega)$, может оказаться эффективным для объектов, которые имеют *гладкие* спектры. Одним из классов таких объектов $o(x)$ являются последовательности дельта-функций, разнесенных на небольшие расстояния, как в рассмотренном примере.

Как негативное следствие изложенного следует указать, что объект, содержащий большое количество импульсов с небольшими интервалами (например, в виде гребенчатой функции Дирака), не может быть эффективно реставрирован с помощью метода экстраполяции, поскольку его спектр сам подобен гребенчатой функции и, следовательно, отнюдь не гладок. По-видимому, этот вывод практически согласуется с обсуждаемыми ниже опытными данными.

5.13. Центральная роль априорного знания

Существует старая поговорка — «экстраполяция опасна». Ее смысл хорошо понимали те, кто брался за решение обсуждаемой проблемы. Однако искушения, вызываемые такими привлекательными вещами, как докторские степени, правительственные контракты и т. п., привели к тому, что возобладала другая поговорка — «было бы желание, а способ найдется»; и действительно, были разработаны некоторые полезные схемы экстраполяции.

Самые эффективные из используемых схем экстраполяции включают в себя наибольшее количество априорных ограничений, накладываемых на выход $\hat{o}(x)$. (Отметим, что индекс «экст» начиная с этого раздела опускается.) Эти ограничения описывают свойства $o(x)$, которые существуют независимо от имеющихся данных $i(y_m)$ и, следовательно, должны независимо вводиться в результат.

5.13.1. Знание ограниченной пространственной протяженности

Важным ограничением указанного типа является знание того, что $o(x)$ отличается от нуля только на конечном интервале $|x| \leqslant X$; см. (5.1). Как ни удивительно, но одно лишь это ограничение обеспечивает однозначное продолжение $\hat{O}(\omega)$ при условии, что в данных об изображении $\{i(y_m)\}$ не имеется шума. В кругах оптиков это положение было высказано, по-видимому, Волтером [44]. Рассмотрим связанные с этим соображения.

Если $o(x) = 0$ при $|x| > X$, то его спектр имеет вид

$$O(\omega) = (2\pi)^{-1/2} \int_{-X}^{X} dx\, o(x) \exp(-j\omega x). \tag{5.81}$$

При этом n-я производная $O^{(n)}(\omega)$ может быть выражена как

$$O^{(n)}(\omega) = (2\pi)^{-1/2} (-j)^n \int_{-X}^{X} dx\, x^n o(x) \exp(-j\omega x).$$

Для любого физически реального $o(x)$ интеграл в правой части должен быть единственным и (вследствие конечности пределов) конечным. Поскольку производные вполне определены, $O(\omega)$ может быть представлен рядом Тейлора относительно любой частоты ω_0. Это в свою очередь означает, что $O(\omega)$ является аналитической функцией.

Аналитическая функция $O(\omega)$ может быть однозначно определена для всех значений ω, если известны 1) все ее производные на одной частоте или 2) ее значения $O(\omega)$ на конечном интервале частот (фактически первое вытекает из второго). Но благодаря инверсной фильтрации, согласно (5.35) и (5.36), оценка $\hat{O}(\omega)$ спектра $O(\omega)$ оказывается известной фактически на всем конечном интервале частот $|\omega| \leqslant \Omega_p$. Следовательно, с помощью инверсной фильтрации образуются все входы, необходимые для экстраполяции информации в частотной области.

По существу были предложены две схемы реставрации с экстраполяцией полосы частот, основанные исключительно на ограниченной пространственной протяженности. Мы дадим здесь краткое описание этих схем, учитывая в основном получаемое с их помощью ясное представление о проблеме экстраполяции. Мы не ожидаем, что они окажутся полезными на практике. Главной проблемой является проникновение шума на выход [согласно разделу 5.12, это связано с недостаточной гладкостью продолжения спектральной функции $O(\omega)$]. Основная причина этого дефекта состоит в том, что единственное априорное ограничение оказывается недостаточно сильным.

Метод Гаррисса

Выражение (5.81) констатирует, что $O(\omega)$ есть функция, ограниченная по полосе пространства. Следовательно, для $O(\omega)$ верна теорема отсчетов (5.7а), которая в данном случае имеет вид

$$O(\omega) = \sum_{n=-\infty}^{\infty} O(n\pi/X)\,\mathrm{sinc}(X\omega - n\pi) \quad \text{для всех } \omega. \tag{5.82а}$$

Благодаря инверсной фильтрации левая часть в принципе известна в пределах оптической полосы частот $|\omega| \leqslant \Omega$. Подставляя далее, скажем, $2N$ значений $O(\omega)$, известного в пределах оптической полосы частот, в (5.82а), мы составим $2N$ уравнений, линейных относительно неизвестных $O(n\pi/X)$, часть из которых лежит вне оптической полосы частот.

Далее с помощью (5.7б) получаем

$$o(x) = (\pi/X) \sum_{n=-\infty}^{\infty} O(n\pi/X) \exp(-jn\pi x/X), \quad |x| \leqslant X, \tag{5.82б}$$

откуда видно, что решение для $2N$ центральных отсчетов $O(n\pi/X)$ образует $2N$-членный ряд Фурье для неизвестного $o(x)$, который обеспечивает экстраполяцию полосы частот, если $N > X\Omega/\pi$.

Таким образом, требуется лишь выбрать в (5.82а) пределы $n = \pm N$ с достаточно большим N и решить $2N$ линейных уравнений для $2N$ неизвестных $O(n\pi/X)$; причем уравнения составляются на основе $O(\omega)$, известного в пределах $|\omega| \leqslant \Omega$. Идея такой экстраполяции была предложена Гаррисом [45].

Однако решение $2N$ линейных уравнений (5.82а) математически эквивалентно решению $2N$ линейных уравнений (5.3). Отметим также соответствие между экспериментальными данными $\{O(\omega_m)\}$ и ФРТ вида $\mathrm{sinc}(X\omega_m - n\pi)$ в (5.82а) с данными $\{i(y_m)\}$ и ФРТ общего вида $s(y_m; x_n)$ в (5.3). Поэтому описываемый метод имеет те же неустранимые недостатки, что и метод прямой инверсии матриц, описанный в п. 5.3.1. Здесь может понадобиться обращение к методам сглаживания (но теперь уже в спектральной области), таким, как метод Филлипса (раздел 5.4) или Туми (раздел 5.5).

Поскольку гладкий спектр $O(\omega)$ имеет сосредоточенное преобразование Фурье $\hat{o}(x)$, такое сглаживание делает эффективным метод Гарриса для решения задачи разрешения двух точек (где как раз требуется гладкий спектр — см. раздел 5.12). С другой стороны, для объектов общего вида этот метод, по-видимому, не имеет преимуществ по сравнению с непосредственной обработкой уравнения отображения, как описано в разделах 5.4 и 5.5.

Использование экстраполирующей функции окна

Обратившись к методу фильтрации, мы должны сразу же задать вопрос — может ли существовать экстраполирующая функция окна $W(\omega)$? Для ее существования требуется, чтобы произведение $W(\omega)O(\omega)$ на конечном интервале $|\omega| \leqslant \Omega$ действовало так же, как $O(\omega)$ на бесконечном интервале, т. е.

$$\int_{-\Omega}^{\Omega} d\omega\, W(\omega)\, O(\omega) \exp(j\omega x) \approx A \int_{-\infty}^{\infty} d\omega\, O(\omega) \exp(j\omega x). \quad (5.83а)$$

Интеграл слева выражает реставрированный выход $\hat{o}(x)$ с учетом функции окна. Поскольку $o(x)$ имеет ограниченную пространственную протяженность, (5.83а) должно удовлетворяться только при значениях

$$|x| \leqslant X. \quad (5.83б)$$

Постоянная A в (5.83а), по-видимому, будет конечной.

Требование (5.83а) кажется невыполнимым. Однако метод Гарриса (5.82а) и (5.82б) подразумевает, что экстраполированный выход получается с помощью чисто линейной операции над $O(\omega)$ в интервале $|\omega| \leqslant \Omega$. Это обнадеживающий признак, поскольку искомый фильтр $W(\omega)$ выполняет как раз линейную операцию.

Фактически функция $W(\omega)$, удовлетворяющая требованиям (5.83а) и (5.83б), была найдена (см. [46]). Это удалось благодаря тому, что функция $O(\omega)$ является ограниченной по полосе пространства (в соответствии с гипотезой об ограниченной протяженности). Решение имеет вид

$$W_N(\omega) = \sum_{n=0}^{N} (-1)^{n/2} \lambda_n(c)^{-1/2} \psi_n(c, 0) \psi_n(c, 2\omega X/\Omega), \quad (5.84а)$$

где n — четные, $c \equiv 2X\Omega$. Чем больше N, тем лучше удовлетворяется требование (5.83а), однако тем меньше становится множитель A. В пределе при $N \to \infty$ (5.83а) идеально выполняется, но $A \to 0$ (реставрация $\hat{o}(x)$ делается равномерно черной).

В (5.84а) входят функции вытянутого сфероида ψ_n и их собственные числа λ_n; эти функции были разработаны Слепяном и его сотрудниками (см., например, [47]). Сводка их свойств и обзор применений в оптике приводятся в работе [48].

Частный случай $O(\omega) = 1$ при $|\omega| \leqslant \infty$ в (5.83а) приводит к требованию, предъявляемому к $W(\omega)$, чтобы ее ПФ $s_{\text{полн}}(x)$, определяемое посредством (5.53), было дельта-функцией. Фактически, как было найдено в работе [30], $W_N(\omega)$ должна удовлетворять уравнению

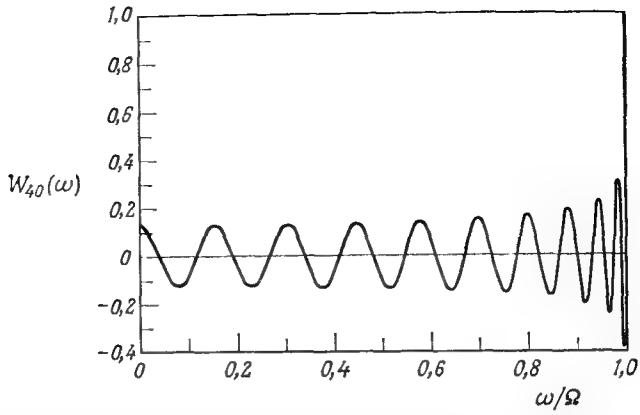
$$s_{\text{полн}}(x) = (2\pi)^{-1/2} \int_{-\Omega}^{\Omega} d\omega W_N(\omega) e^{i\omega x} = A\delta_N(x), \quad |x| \leqslant 2X, \quad (5.84б)$$

где $\delta_N(x)$ — приближение N-го порядка к дельта-функции вида

$$\delta_N(x) = \sum_{n=0}^{N} \lambda_n(c)^{-1} \psi_n(c, 0) \psi_n(c, x), \quad (5.84в)$$

где $|x| \leqslant 2X$, n — четные. Графики $W_N(\omega)$ и $\delta_N(x)$ для случая $N = 40$, $c = 6{,}25$ показаны на фиг. 5.9 и 5.10.

Автор этой главы занялся исследованием функций $W_N(\omega)$ после прочтения статьи Барнса [49], который первым применил функции $\psi_n(\omega)$ для решения проблемы реставрации. Впоследствии для функций $W_N(\omega)$ удалось найти интересное приложение, используя ее в качестве функции пропускания зрачка оптической системы, действующей в реальном масштабе времени [30, 48].



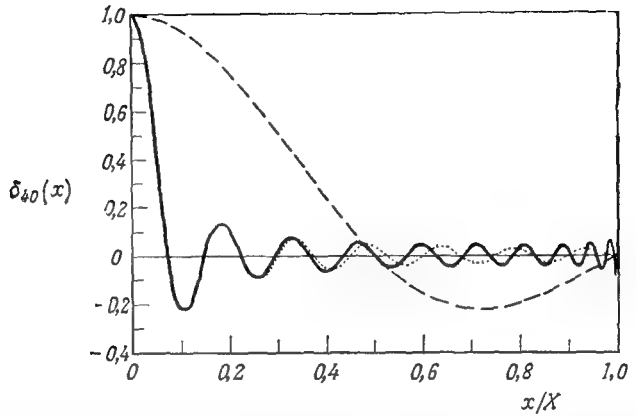
Фиг. 5.9. Экстраполирующая функция окна $W_{40}(\omega)$ для случая $c = 6{,}25$. Кривая симметрична относительно нулевой частоты. В широкой области у нулевой частоты она хорошо аппроксимируется косинусоидой. Частота и амплитуда W_{40} быстро увеличиваются с приближением к граничной частоте Ω [47].

Конечно, при реальном использовании функции окна $W_N(\omega)$ возникает трудность, связанная с тем, что экстраполируются без различия как сигнальная, так и шумовая составляющие входного изображения $I(\omega)$. Поскольку полный фильтр, обрабатывающий $I(\omega)$, выражается в виде $W_N(\omega)/\tau(\omega)$, то полный выход фильтра имеет вид $W_N(\omega)O(\omega) + W_N(\omega)[N(\omega)/\tau(\omega)]$. До сих пор мы рассматривали только первый из этих двух членов выходного интеграла [в левой части (5.83а)].

Второй член обусловливает шумовой выход, который имеет вид

$$\int_{-\Omega}^{\Omega} d\omega W_N(\omega)[N(\omega)/\tau(\omega)]\exp(j\omega x). \qquad (5.84г)$$

Этот член приводит к усугублению трудностей, связанных с возрастанием шума при применении инверсной фильтрации в чистом виде. Поскольку с ростом ω шумовой член $N(\omega)$ сохраняет постоянное значение, а $\tau(\omega) \to 0$, отношение N/τ в (5.83в) возрастает по мере приближения к граничной частоте Ω. Но затем функция W_N производит экстраполяцию этого усиленного шумового члена (для простоты предполагается, что N/τ также является функцией, ограниченной по полосе пространства). Эта экстраполированная часть шума, следовательно, поднимается еще выше в конечной полосе частот ω за граничной частотой $\pm\Omega$ и лишь после этого снова опускается к нулю. Понятно, что это приводит к существенному усилению шума по сравнению с инверсной фильтрацией.



Фиг. 5.10. Сплошной линией показано преобразование Фурье $\delta_{40}(x)$ от функции окна на фиг. 5.9; эта кривая является также графиком функции $s_{полн}(x)$. Для сравнения штриховой линией показана полная функция рассеяния для инверсной фильтрации в полосе частот функции $W_{40}(\omega)$. Пунктирной линией показана полная функция рассеяния для инверсной фильтрации в полосе частот, превышающей полосу $W_{40}(\omega)$ в 6 раз. Таким образом, в результате умножения на функцию окна W_{40} в (5.83а) спектр объекта при выбранном значении c экстраполируется в 6 раз [47].

Хотя были сделаны попытки дополнить этот основной метод некоторыми средствами подавления шума (см. [17, 50, 51]), перспективы его практического применения сомнительны.

Кроме серьезной проблемы экстраполяции шума, общий подход (5.83а) не свободен от таких недостатков, как значительная паразитная осцилляция, появление отрицательных выходов и чисто вычислительная проблема генерирования требуемых функций $\psi_n(\omega)$. Знание только конечной протяженности, очевидно, не может оказать достаточно сильного влияния на оценку $\hat{o}(x)$, чтобы обеспечить значительную экстраполяцию в присутствии шума. Необходимы какие-то дополнительные сведения.

5.13.2. Знание положительности: влияние на паразитную осцилляцию и разрешение

Простое ограничение в виде требования положительности оценки $\hat{o}(x)$ оказывает на нее сильное и благотворное воздействие. Покажем это, развивая соображения, приведенные в п. 5.13.1.

Предположим для простоты, что объект $o(x)$ состоит из одной дельта-функции $\delta(x)$. Тогда изображение принимает форму дискретных отсчетов функции рассеяния $s(y_m)$, которые практически равны нулю везде, кроме центральной области $|y_m| \leqslant y_0$.

Рассмотрим точки объекта, сопряженные с этими весьма ма-

лыми отсчетами изображения. Допустим, что реставрируемый объект должен быть ограничен только положительными значениями. Тогда, если в этих точках возникает паразитная осцилляция, она должна происходить относительно некоторого положительного локального среднего. Следовательно, суммарные вклады от этих осцилляций в интеграл изображения (5.1) также должны быть положительными и конечными. Однако в соответствии с принятой гипотезой отсчеты входного изображения в этой области $|y_m| > y_0$ пренебрежимо малы. Поэтому подобные осцилляции объекта оказываются несовместимыми с данным изображением. Вследствие этого объект, ограниченный положительными значениями и согласованный с входными данными об изображении, не может иметь паразитных осцилляций в областях, с которыми сопряжены области истинного объекта, где его значения пренебрежимо малы.

Описанные явления имеют преобладающее значение для объектов, состоящих главным образом из нулевых значений, т. е. содержащих конечное число дельта-функций на нулевом фоне. В то же время было эмпирически установлено, что отсутствие паразитной осцилляции часто сопровождается увеличенным разрешением выхода $\hat{o}(x)$. Это можно объяснить, используя рассуждения, подобные приведенным в п. 5.13.1.

Действительно, если для реставрируемого объекта допускается заход в отрицательную область, т. е. не вводится ограничение по положительности, то типичная реставрация объекта $o(x) = \delta(x)$ будет иметь вид $\operatorname{sinc}(\Omega x)$. В ней присутствует паразитная осцилляция, начинающаяся с большого отрицательного лепестка при $x = \pi/\Omega$. Очевидно, что вклад этого отрицательного лепестка в изображение компенсируется за счет соседних точек в положительном центральном лепестке $\operatorname{sinc}(\Omega x)$. Следовательно, если отрицательный лепесток не существует (благодаря ограничению реставрации положительными значениями), то соседние точки положительного лепестка уже не должны быть положительными и конечными. Они могут оставаться равными нулю, что приводит к сужению центральной области и тем самым к увеличению разрешения.

5.13.3. Знание верхней границы

Глядя на изображение, наблюдатель всегда имеет возможность сделать предположение относительно некоторой верхней границы B неизвестного объекта. (Заметим, что B может не быть точной верхней гранью, т. е. числом, действительно достигаемым $\hat{o}$ при некотором x.) Таким образом, пользователь получает возможность ограничить $\hat{o}(x)$ пределами

$$0 \leqslant o(x) \leqslant B. \tag{5.85}$$

Джанссон [10] установил эмпирически, насколько важна эта дополнительная информация. В качестве объектов он рассматривал спектры поглощения, которые должны быть ограничены по уровню значениями 0 и 100%. Ниже мы подробно описываем метод реставрации Джанссона.

Объяснение, почему важно знание верхней границы, можно дать, используя аргументы, совершенно аналогичные рассмотренным в п. 5.13.2 относительно нижней границы. Эти два случая вполне симметричны. Поэтому мы можем сделать вывод, что информация о верхней границе полезна в основном для областей, соответствующих уровню B в истинном объекте. Это означает, что ограничение сверху будет особенно эффективно, когда верхняя граница, предположительно устанавливаемая пользователем, окажется точной верхней гранью, т. е. тем уровнем, который действительно достигается в различных точках объекта (и чем в большем числе точек, тем лучше).

Что касается информации о наличии комбинированного ограничения снизу и сверху, то, как следует из сказанного выше, ее можно использовать наилучшим образом при реставрации объектов, которые целиком состоят из частей, соответствующих либо верхней, либо нижней известной границе. В случае такого двоичного объекта алгоритм реставрации сводится к простой оценке положения границ в объекте. Объектами такого типа являются составные части спектров поглощения и буквенно-цифровая информация.

5.14. Общие замечания о методах реставрации с ограничением по положительности

Если читатель уже убедился в важности требования положительности при реставрации $\hat{o}(x)$, он вправе спросить, как обеспечить его выполнение. Точнее, как можно сформировать согласованное с входными данными об изображении $i(y_m)$ и положительное $\hat{o}(x)$ посредством свертки (5.1)? Это важнейший вопрос. Если, например, положительное $\hat{o}(x)$ формируется просто путем взятия абсолютного значения оценки без ограничения $\hat{o}(x)$, то такое положительное $\hat{o}(x)$ уже не будет согласовано с данными об изображении. В нем не удастся обнаружить улучшения разрешения или гладкости, поскольку в рассуждениях п. 5.13.2 всегда подразумевалась такая согласованность.

Прежде всего следует заметить, что никакой линейный метод (типа свертки) не может дать гарантированного положительного выхода при всех x. Это следует из того, что для улучшения разрешения ядро свертки должно иметь отрицательные выбросы (в противном случае свертка приведет просто к дальнейшему сглаживанию, или смазыванию, изображения). Эти отрицатель-

ные выбросы будут сказываться на выходе в областях, соответствующих нулевому фону истинного объекта.

По этой причине приходится обратиться к нелинейным операциям над данными об изображении $\{i(y_m)\}$. В эту категорию попадают схемы типа $\hat{o}(x)=[i(x)]^2$, однако такие $\hat{o}(x)$, очевидно, не будут согласованными с данными об изображении. Тем не менее заметим между прочим, что если $i(x)$ ограничено полосой частот 2Ω, то подобное $\hat{o}(x)$ будет иметь полосу частот 4Ω, т. е. вдвое шире входной. Такая тенденция к расширению полосы частот свойственна всем нелинейным методам.

Итак, проблема состоит в том, чтобы найти такую нелинейную функцию от $i(y)$, которая гарантирует а) положительность и б) согласованность с изображением. Для выполнения этого двойного требования разработано несколько алгоритмов. Они описываются в следующих разделах. В частности, будет показано, что для удовлетворения второго требования все разработанные методы (кроме одного) используют итеративный поиск решения. Поэтому разработанные методы в общем требуют большего времени вычислений, чем описанные в предыдущих разделах линейные методы.

5.15. Метод Шелла и Биро

Очевидный путь обеспечения положительного выхода $o(x)^+$ состоит в том, чтобы принять

$$o(x)^+ = a(x)^2 \tag{5.85а}$$

и затем попытаться найти $a(x)$. В частотной области, используя тождество (5.9), имеем

$$O(\omega)^+ = A(\omega) \otimes A(\omega). \tag{5.85б}$$

$A(\omega)$ можно определить, исходя из требования согласованности с изображением. Согласованность означает, что в полосе частот данных $|\omega| \leqslant \Omega_p$ должно выполняться

$$I(\omega) = \tau(\omega)[A(\omega) \otimes A(\omega)]. \tag{5.86а}$$

Это условие можно также записать в виде

$$I(\omega)/\tau(\omega) = A(\omega) \otimes A(\omega), \quad |\omega| \leqslant \Omega_p. \tag{5.86б}$$

Однако левая часть (5.86б) соответствует оценке с помощью инверсной фильтрации, полученной в разделе 5.9. Мы видим, что в качестве входов для алгоритма положительного ограничения можно использовать дискретные отсчеты решения $O(\omega_m)$, полученного с помощью инверсной фильтрации.

Итак, согласованность выражается посредством (5.86б). Однако из-за шума в измеряемом изображении выходы $\hat{O}(\omega_m)$

могут содержать ошибку. Поэтому прежде, чем потребовать выполнения строгого равенства (5.86б), мы можем использовать более слабое условие

$$E \equiv \sum_{\omega_m=-\Omega_p}^{\Omega_p} |\hat{O}(\omega_m) - A(\omega_m) \otimes A(\omega_m)|^2 = \min, \qquad (5.87а)$$

выполнение которого зависит от выбора $A(\omega)$.

С другой стороны, на частоте $\omega = 0$, где $\tau(\omega) = 1$, ошибка инверсной фильтрации минимальна [см. (5.34)]. Поэтому для данной точки мы сохраняем требование строгого равенства:

$$\hat{O}(0) = A(\omega) \otimes A(\omega)\,|_{\omega=0}. \qquad (5.87б)$$

В выражениях (5.87а) и (5.87б) содержится формулировка задачи, данная Биро.

Необходимо сделать еще некоторые замечания относительно требования (5.87б). Две части (5.87б) представляют собой площади под кривыми $\hat{o}(x)$ и $o(x)^+$. Требование равенства означает, что окончательный выход $o(x)^+$ должен получаться в результате чистого перераспределения энергии (без усиления) под ограниченной по полосе входной кривой $\hat{o}(x)$. Кроме того, поскольку $\hat{o}(x)$ линейно связано с истинным $o(x)$, выражение (5.87б) устанавливает полную энергию выхода $o(x)^+$ равной энергии истинного объекта (с добавлением некоторого шума). Поскольку эта величина близка к достоверной оценке, можно было ожидать, что она поможет получению для (5.87а) сходящегося решения. Это предположение получило эмпирическое подтверждение.

Здесь важно заметить, что аппроксимация $\hat{O}(\omega)$ в (5.87а) и (5.87б) в пределах $|\omega| \leqslant \Omega_p$ требует определения $A(\omega)$ только на полуинтервале $|\omega| \leqslant \Omega_p/2$. Таково свойство операции свертки в (5.87а). Однако если для частот $\omega_m > \Omega_p/2$ $A(\omega) \neq 0$, то эти значения также войдут в (5.87а) и (5.87б); они «замешиваются» в интервал $|\omega| \leqslant \Omega_p$ посредством операции свертки. В этом и состоит способ укладки большего числа степеней свободы A в интервал данных $|\omega| < \Omega_p$ с целью дальнейшего уменьшения минимума, достигаемого в (5.87а). Ниже мы еще вернемся к этому вопросу.

Предположим, что относительно некоторых составляющих $\hat{O}(\omega_m)$ известно, что они с большей вероятностью подвержены ошибкам, чем другие составляющие (в соответствии с имеющимися априорными сведениями об энергетических спектрах сигнала и шума). При этом (5.87а) полезно видоизменить путем введения мультипликативных весов для каждой ω_m так, чтобы уменьшить вероятное значение минимума (см. [8]). Укажем для примера, что малый вес при ω_k дает возможность свертке

$A \otimes A(\omega_k)$ значительно отклоняться от $\hat{O}(\omega_k)$, не затрагивая существенно соответствующего вклада в минимум. Такой прием логичен, если известно, что при ω_k отношение сигнала к шуму мало. Однако без знания статистических характеристик сигнала и шума этот способ оказывается неосуществимым.

Шелл и Биро решили задачу, выражаемую (5.87а) и (5.87б), различными способами, но оба использовали итеративные алгоритмы. Алгоритм Шелла имеет вид

$$A^{(k+1)}(\omega_m) = C \sum_{n=-(M+m)/2}^{(M-m)/2} [\hat{O}(\omega_n) - A^{(k)} \otimes A^{(k)}(\omega_n)] \cdot A^{(k)}(\omega_{n+m}), \quad (5.88)$$

где C — постоянная, $k = 0, 1, 2, \ldots$. Индекс (k) обозначает номер итеративной оценки $A(\omega_m)$; $m = 0, \pm 1, \pm 2, \ldots, \pm P$; $P \geqslant M/2$; $2M$ — полное число входных отсчетов $\hat{O}(\omega_n)$. Начальная оценка $\{A^{(0)}(\omega_n)\}$ подставляется в правую часть; при этом в левой части получается $\{A^{(1)}(\omega_n)\}$. Затем в правую часть подставляется $\{A^{(1)}(\omega_n)\}$ и т. д., пока не будет выполнено определенное число итераций $k = K$, для которого $\{A^{(K+1)}(\omega_n)\} = \{A^{(K)}(\omega_n)\} +$ допустимая ошибка. Такая вычислительная процедура называется методом последовательной подстановки (см., например, [52]).

Алгоритм (5.88) может быть получен путем внесения (5.87б) в (5.87а) в качестве аддитивного условия Лагранжа и приравнивания нулю производной $\partial/\partial A(\omega_k)$ полученного выражения. Левая часть (5.88) по существу соответствует производной ограниченного члена (5.87б) в составе этого выражения. Использованная в (5.88) нумерация итераций $(k+1)$, (k) обусловлена методом последовательной подстановки.

Способ решения уравнений (5.87а) и (5.87б), разработанный Биро, в общих чертах описан в работе [9]. Он основан на последовательных возмущениях неизвестных $A(\omega_n)$, определяемых как

$$A(\omega_n) = A(n\Delta\omega), \qquad n = 0, \pm 1, \pm 2, \ldots, \pm P. \quad (5.89а)$$

Входами служат

$$\hat{O}(\omega_m) = \hat{O}(m\Delta\omega), \quad m = 0, \pm 1, \pm 2, \ldots, \pm M, \quad (5.89б)$$

где

$$M\Delta\omega = \Omega$$

устанавливает интервал данных в частотной области. Шаг $\Delta\omega$ определяется как интервал дискретизации с учетом (5.7б) и известной конечной протяженности $|x| \leqslant X$ исходного объекта и составляет

$$\Delta\omega = \pi/X. \quad (5.89в)$$

Для первой из описываемых ниже серий возмущений $P = M/2$. В этом случае отсчетов $A(\omega_n)$ как раз достаточно для заполнения (5.87а) во всем интервале данных $|m| \leqslant M$.

Вначале за неизвестные в (5.87а) и (5.87б) принимаются три величины $A(\omega_0)$, $A(\omega_1)$ и $A(\omega_{-1})$, тогда как все остальные A берутся равными нулю. Эти три неизвестные определяются как решения трех кубических уравнений, получаемых следующим образом. Присоединим условие (5.87б) к (5.87а) в качестве аддитивного лагранжева члена. Затем приравняем нулю производные $\partial/\partial A(\omega_0)$, $\partial/\partial A(\omega_1)$ и $\partial/\partial A(\omega_{-1})$. Полученные три кубических уравнения решаются с помощью эффективного алгоритма.

После выполнения первой аппроксимации $A(\omega_0)$, $A(\omega_1)$ и $A(\omega_{-1})$ обращаемся к поиску $A(\omega_0)$, $A(\omega_2)$ и $A(\omega_{-2})$. Первое решение для $A(\omega_{\pm 1})$ при этом сохраняется, а остальные A приравниваются нулю. Затем вновь составляются и решаются три кубических уравнения для трех рассматриваемых неизвестных.

Далее обращаются к поиску $A(\omega_0)$ и $A(\omega_{\pm 3})$. При этом ранее полученные значения $A(\omega_{\pm 1})$ и $A(\omega_{\pm 2})$ сохраняются, а остальные A приравниваются нулю. Три новых неизвестных $A(\omega_0)$, $A(\omega_{\pm 3})$ находятся, как описано выше. Эта процедура продолжается до тех пор, пока не будет получена тройка неизвестных $A(\omega_0)$, $A(\omega_{\pm P})$.

Такие возмущения неизвестных, которые производятся поочередно по тройкам до получения $A(\omega_{\pm P})$, образуют один цикл процедуры возмущения.

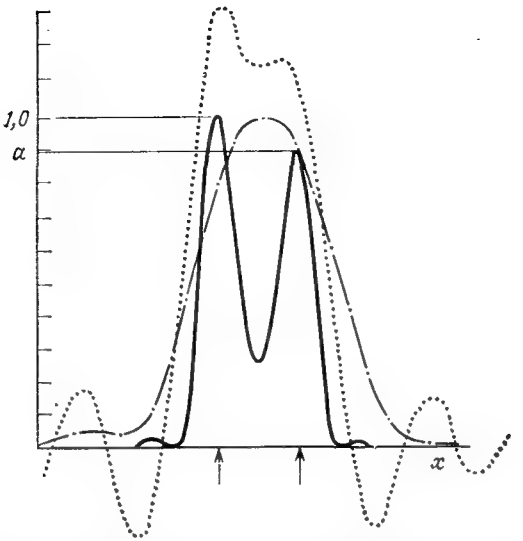
После этого новый цикл начинается с того, что в качестве неизвестных берут опять центральную тройку $A(\omega_0)$, $A(\omega_{\pm 1})$, а для всех остальных A сохраняют прежние значения; затем переходят к частотам $\omega_{\pm 2}$ и т. д. После того как будет вновь определена тройка неизвестных $A(\omega_0)$, $A(\omega_{\pm P})$, заканчивается второй цикл.

Результат каждого нового цикла k выражается в дальнейшем уменьшении минимума $E(k)$ в (5.87а). Однако после конечного числа циклов, скажем K, обнаруживается «насыщение», т. е. минимум перестает сколько-нибудь существенно уменьшаться. Что же делать дальше? И главное следует ли продолжать минимизацию?

Заметим, что входы $\hat{O}(\omega_m)$ образованы в результате инверсной фильтрации и поэтому содержат шум $N(\omega_m)/\tau(\omega_m)$ [см. (5.34)]. Подстановка $O + N/\tau$ вместо $\hat{O}$ в (5.87а) фактически показывает, что при условии $A \otimes A = 0$ (т. е. при идеально точной оценке A) образуется остаточный ненулевой минимум

$$\sum_{m=-M}^{M} |N(\omega_m)/\tau(\omega_m)|^2 = E_{\text{ост}}. \tag{5.90}$$

Пользователь может узнать вероятное значение $\langle E_{\text{ост}} \rangle$, располагая сведениями об энергетическом спектре шума и функции передачи системы.



Фиг. 5.11. Различные варианты распределения яркости звезды альфа Лебедя, полученного с помощью радиотелескопа в Нанси. Штрих-пунктиром показана запись измерения, пунктиром — результат инверсной фильтрации этой записи, а сплошной линией — положительная реставрация по методу Биро. Разделение двух пиков и отношение их яркостей a довольно хорошо согласуются с результатами прямого измерения телескопом с большей апертурой [9].

Поэтому логично продолжать попытки уменьшить минимум $E^{(k)}$ в (5.87а) только при условии, что он превышает значение $\langle E_{\text{ост}}\rangle$. Допустим, что это так и есть, но при этом, как упоминалось выше, $E^{(k)}$ уже доведено до неуменьшающейся величины. В этом случае у пользователя остается единственный резерв — знание того, что разрешенные значения $A(\omega_{P+1})$, $A(\omega_{-P-1})$, если они отличны от нуля, могут добавить в сумму (5.87а) новые степени свободы и тем самым обеспечить дальнейшее снижение E.

Учитывая это, далее производят возмущение тройки $A(\omega_0)$, $A(\omega_{P+1})$, $A(\omega_{-P-1})$, оставляя для всех остальных A прежние значения. Обычно при этом происходит весьма заметное уменьшение E. Затем снова производят возмущение самой центральной тройки, и этот процесс продолжают как прежде, используя, однако, в качестве наивысших частот ω_{P+1}, ω_{-P-1}.

Если для E вновь наступает насыщение и если при этом $E > \langle E_{\text{ост}}\rangle$, то охватывают возмущениями следующие более высокие значения частот ω_{P+2}, ω_{-P-2} и т. д. В конце концов E становится ниже $\langle E_{\text{ост}}\rangle$ и мы получаем искомое решение $o(x)^+$ в виде

$$o(x)^+ = \sum_{n=-Q}^{Q} O(\omega_n)^+ e^{i\omega_n x}, \qquad (5.91)$$

где $O(\omega_n)^+$ определяется выражением (5.85б). Здесь ω_Q — наивысшая частота, использованная в процессе возмущений. Вообще говоря, ω_Q превышает полосу частот данных Ω. Следовательно, в таких случаях производится экстраполяция полосы частот.

Эффективной демонстрацией способа Биро было его применение для определения яркостного профиля двойной звезды альфа Лебедя. Экспериментальные данные были получены с помощью радиотелескопа в Нанси; они имели отношение сигнал/шум около 50. Соответствующие кривые показаны на фиг. 5.11.

Подробное исследование алгоритма Биро составило предмет диссертации Вонга [53].

5.15.1. Обсуждение

Для метода Шелла и Биро характерно, что при необходимости получить положительный выход, согласованный с входными данными, приходится выходить за пределы полосы частот данных. При этом предполагается, что остаточная ошибка $E_{\text{ост}}$, учитывающая усиление шума, согласно (5.90), достаточно мала. Практически метод допускает довольно существенные уровни шума на входе, а двукратная и более сильная экстраполяция является скорее правилом, чем исключением.

Однако преимущества этих методов проявляются только на объектах, которые состоят из произвольной решетки импульсов на известном (например, нулевом) фоне. Поэтому, например, объект, состоящий из случайных ступенек, реставрируется обсуждаемым методом не намного лучше, чем с помощью плоской инверсной фильтрации. Граничные переходы улучшаются при этом сильнее, чем при инверсной фильтрации, но возникает большая осцилляция на плоских участках. Такой недостаток типичен для всех существующих схем реставрации с ограничением по положительности. По этому поводу см. также п. 5.19.2.

5.16. Метод Джанссона — ван Циттерта

В п. 5.10.4 мы исследовали метод последовательной свертки ван Циттерта и выяснили, почему он обычно не пригоден для резкого улучшения разрешения зашумленных данных. Недавно Джанссон [10] произвел модификацию этого метода, превратившую его в полезный инструмент улучшения спектральных данных.

Джанссон модифицировал первоначальный алгоритм ван Циттерта (5.59) следующим образом. Исходим из тех же начальных условий $k = 0$, $\hat{o}^{(0)}(x) = i(x)$. Затем производим итерации:

$$1)\ i^{(k)}(x) = \hat{o}^{(k)}(x) \otimes s(x),$$

$$2)\ \hat{o}^{(k+1)}(x) = \hat{o}^{(k)}(x) + r(x)[i(x) - \hat{i}^{(k)}(x)],$$

где

$$r(x) = C\left[1 - 2\left|\hat{o}^{(k)}(x) - \frac{1}{2}\right|\right], \quad C \text{ — постоянная,}$$

3) $k \to k+1$, (5.92)

4) перейти к (1).

Фактически $\hat{o}^{(k)}(x)$ в правой части шага (1) включает в себя имеющиеся на этот момент значения $o^{(k-1)}(x)$ из шага (2), поскольку они уже сформированы при предшествующих значениях k; но ради упрощения записи это не учитывается в обозначениях.

Сравнивая этот метод с методом ван Циттерта, мы замечаем, что на шаге (2) модифицированного алгоритма вводится коэффициент релаксации r. В этом и состоит сущность модификации. Зависимость $r(x)$ от $\hat{o}^{(k)}(x)$ такова, что $r(x) \to 0$, когда $\hat{o}^{(k)}(x)$ стремится к 0 или к 1. Поэтому если при любом x итерация $\hat{o}^{(k)}$ на шаге (2) оказывается близкой к 0 или 1, то следующая итерация $\hat{o}^{(k+1)}$ и все остальные итерации сохраняют полученное значение. Это ограничивает выход $\hat{o}(x)$ пределами $0 \leqslant \hat{o}(x) \leqslant 1$. Именно такому физическому требованию удовлетворяют спектры поглощения, для которых был разработан этот алгоритм.

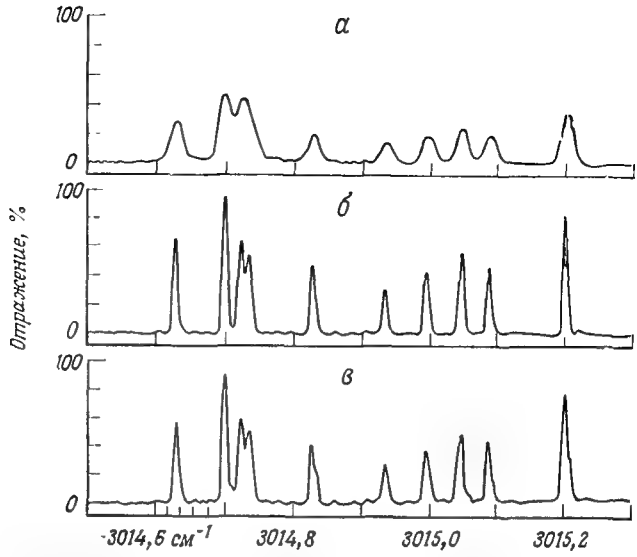
Возвращаясь к шагу (2), мы видим, что вначале r линейно относительно $\hat{o}^{(0)}$, которое в свою очередь линейно относительно изображения i. Но затем $\hat{o}^{(1)}$, формируемое на шаге (2), становится квадратичным относительно изображения i, $\hat{o}^{(2)}$ становится кубичным относительно i и т. д. В общем выход $\hat{o}(x)$ оказывается весьма нелинейно связанным с данными об изображении. Таким образом, коэффициент $r(x)$ превращает линейный и не учитывающий ограничений метод ван Циттерта в нелинейный и учитывающий ограничения метод.

Если описанный метод понадобится для обработки изображений другого типа, чем спектры поглощения, то $r(x)$ можно приспособить к выполнению новых требований по ограничению. Если желательно, например, чтобы $A \leqslant \hat{o}(x) \leqslant B$, где A и B — известные постоянные, то можно использовать новый коэффициент релаксации в виде

$$r(x) = C\left[1 - 2(B-A)^{-1}\left|\hat{o}^{(k)}(x) - 2^{-1}(A+B)\right|\right].$$

Если же требуется просто положительный выход $\hat{o}(x) \geqslant 0$, то в предыдущем выражении можно установить $A = 0$ и использовать любое B, которое ограничивает сверху ожидаемую картину $o(x)$.

Более интересно, что описанную процедуру можно обобщить на случай, когда пользователь знает ограничивающие кривые $A(x)$ и $B(x)$. При этом постоянные A и B в предыдущей формуле заменяются функциями $A(x)$ и $B(x)$.



Фиг. 5.12. Иллюстрация применения метода Джанссона для реставрации спектроскопических данных. *а* — исходный спектр, соответствующий ветви Q полосы ν_3 для CH_4. *б* — результат реставрации с использованием измеренной ФРТ. *в* — результат реставрации, когда измеренная ФРТ моделировалась гауссовой функцией, ширина которой на половине высоты была такой же, как у измеренной ФРТ. Такая аппроксимация вызывает малозаметное снижение качества реставрации [10].

На фиг. 5.12 показан типичный пример использования алгоритма (5.92) для реставрации спектроскопического изображения. Разрешение на графиках *б* и *в* значительно выше, чем на графике *а*, так что одна из спектральных линий даже расщепляется на две. График *в* демонстрирует, что обсуждаемый метод не слишком чувствителен к точности формы функции рассеяния $s(x)$; в данном случае в алгоритме (5.92) была использована гауссова аппроксимация функции рассеяния, полученной экспериментально.

5.17. Реставрация по максимуму энтропии: две идеи, два подхода

Алгоритмы реставрации с ограничениями, описанные в разделах 5.15 и 5.16, в сущности были разработаны для решения частных задач. Поэтому оба алгоритма выражают прагматическую последовательность операций на основе интуиции. Уместно задать вопрос, имеется ли для решения подобных задач надеж-

ный формальный подход, логически вытекающий из какого-либо известного разумного принципа (в качестве примера укажем метод фильтрации Винера — Хелстрома, построенный на основе критерия МСКО; см. п. 5.11.1). К этой категории принадлежит и принцип максимума энтропии.

Рассмотрим основания к его применению. Задача инверсии (5.1) очень близка к задаче инверсии выражения

$$q_m = \int_{-\infty}^{\infty} dx\, x^m \hat{p}(x), \quad m = 1, 2, \ldots, M, \tag{5.93а}$$

для определения неизвестной плотности вероятности $\hat{p}(x)$ по M известным ее моментам $\{q_m\}$. Моменты здесь соответствуют входным данным об изображении $\{i(y_m)\}$. Это подобие простирается еще дальше, поскольку, чтобы представлять реальную плотность вероятности, $\hat{p}(x)$ должна быть положительной. Джейнис [11] показал, что при решении этой задачи наименее смещенную оценку $\hat{p}(x)$ дает только та функция, которая имеет максимум энтропии

$$H = -\int_{-\infty}^{\infty} dx\, \hat{p}(x) \ln \hat{p}(x) \tag{5.93б}$$

и соответствует приведенным выше уравнениям моментов. «Наименее смещенная» — значит равновероятная оценка в той степени, в какой это допускают входные моменты; тем самым условие максимума энтропии оказывает на оценку сглаживающее воздействие. Начиная с п. 5.3.1, мы имели возможность увидеть, насколько важным является этот эффект. Однако обсуждаемый подход имеет еще одно достоинство.

Требуемый, согласно методу Джейниса, максимум энтропии обеспечивается решением для $\hat{p}(x)$ в виде

$$\hat{p}(x) = \exp\left[-\sum_{m=1}^{M} \lambda_m x^m\right], \tag{5.93в}$$

где неизвестные λ_m должны быть взяты такими, чтобы удовлетворить уравнениям моментов. При такой форме выражения $\hat{p}(x)$ не может быть отрицательной. Вследствие этого критерий максимума энтропии автоматически обеспечивает чисто положительный выход.

Мы убедились в полезности такого подхода и хотели бы узнать, как его применить для оценки пространственной (а не вероятностной) функции. Очевидно, o (при любом x) следует теперь моделировать посредством случайной переменной так, чтобы можно было сформировать ее вероятность $p(o)$. Когда $p(o)$ известна, энтропия H определяется с помощью (5.93б).

С этого момента начинается расхождение между методами Бурга [5] и Фридена [18].

Согласно методу Бурга, o рассматривается как квадрат другой переменной, обозначаемой a (как в разделе 5.15). Этим обеспечивается положительность o. Далее первые два момента величины a предполагаются известными (по крайней мере для оптических задач такое предположение является произвольным). Из (5.93в) с учетом этого предположения следует, что наименее смещенным представлением $p(a)$ будет нормальное (а в предположении нулевого среднего — гауссово) распределение.

При таких условиях энтропия оказывается известной. Для точки x она составляет

$$H = -\alpha \ln \langle o \rangle + \beta,$$

где α и β — действительные числа. При переходе ко всей оси x наименьшую ошибку дает предположение о статистической независимости $o(x)$. При этом общая энтропия будет просто суммой отдельных энтропий, что дает

$$H = -\alpha \int dx \ln \langle o(x) \rangle + \beta.$$

Последнее, что еще нужно сделать, — это принять, что $\langle o(x) \rangle$ представляет собой искомую оценку $\hat{o}(x)$. И, наконец, мы получаем критерий

$$H = -\int dx \ln \hat{o}(x) = \max, \tag{5.94}$$

причем коэффициенты α и β не имеют значения при определении $\hat{o}(x)$, удовлетворяющего (5.94). В этом состоит метод максимума энтропии, приспособленный к решению оптических задач.

Другое применение принципа максимума энтропии было независимо разработано Фриденом [18]. Принятая им статистическая модель объекта коренным образом отличается от модели Бурга. В данном случае объект считается состоящим из дискретных математических «зерен» малой интенсивности Δo, которые распределены по всему полю объекта. Поле поделено на клетки, сцентрированные относительно подмножества точек $\{x_n\}$; при этом предполагается, что неизвестный объект имеет O_n зерен в клетке n. Таким образом, $o(x_n) = O_n \Delta o$. Пусть p_n представляет вероятность появления зерна в клетке n. Тогда при условии, что по объекту распределено большое количество зерен, согласно закону больших чисел, имеем

$$p_n = O_n / O_T,$$

где O_T — общее число зерен в объекте. O_T считается известным на основе сохранения энергии в данных об изображении.

С учетом введенных обозначений энтропия выражается в виде

$$H \equiv -\sum_{n=1}^{N} p_n \ln p_n = -\sum_n (O_n/O_T) \ln (O_n/O_T) =$$
$$= -\sum_n [o(x_n)/\Delta o O_T] \ln [o(x_n)/\Delta o O_T].$$

Используя определение O_T, окончательно получаем

$$H = -\alpha \sum_n o(x_n) \ln o(x_n) + \beta,$$

где α и β — постоянные. В результате принцип максимума энтропии принимает вид

$$H = -\sum_n \hat{o}(x_n) \ln \hat{o}(x_n) = \max. \qquad (5.95)$$

Сравнение этого критерия с (5.94) показывает, что критерии различаются множителем $o(x)$. Этот множитель придает критерию (5.95) форму, более близкую к классической энтропии (5.93б), чем у критерия (5.94). Однако за это уподобление приходится платить, как будет показано ниже, потерей возможности выразить решение для $\hat{o}(x)$ в замкнутой форме. Наконец, следует отметить, что (5.95) выведено с использованием, по-видимому, более скромного набора допущений, чем (5.94).

Важно, однако, что оба критерия оказывают сглаживающее воздействие на оценку $\hat{o}(x)$. Действительно, когда два значения объекта $o(x_1)$ и $o(x_2)$ вынуждают дифференциально изменяться так, чтобы они стали более одинаковыми (что и выражает тенденцию к сглаживанию), оба варианта H возрастают. Следовательно, максимум H, согласно обоим критериям, способствует получению максимально гладкой оценки, а абсолютный максимум H образуется при идеально гладкой оценке $\hat{o}(x) = \text{const}$. Однако входные данные выступают при этом как ограничения, которые вызывают флуктуации в $\hat{o}(x)$ и тем самым нарушают ситуацию, в которой H имеет абсолютный максимум.

5.18. Решение Бурга для максимума энтропии

Как и в методе Шелла и Бриро, предположим, что в качестве входных данных используются оценки с помощью инверсной фильтрации $\{\hat{O}(\omega_m)\}$. Далее эти оценки служат для того, чтобы обычным образом ввести ограничения в энтропийный критерий

(5.94), назначая множители Лагранжа λ_m для каждого выражения

$$-\int_{-X}^{X} dx \ln \hat{o}(x) - \\ -\sum_{m=1}^{M} \lambda_m \left[\int_{-X}^{X} dx\, \hat{o}(x)\, e^{-i\omega_m x} - \hat{O}(\omega_m) \right] = \max. \quad (5.96а)$$

Вывод был проделан (но не опубликован) Бургом и лишь недавно появился в открытой литературе [54]. Поскольку он довольно велик, мы его не приводим. Результат имеет вид

$$\hat{o}(x) = \frac{(2X)^{-1} \gamma_{M+1}}{\left| 1 + \sum_{m=1}^{M} \gamma_m \exp\left[-jm\pi (2X)^{-1} (x + X) \right] \right|^2}, \quad (5.96б)$$

причем $(M+1)$ неизвестных $\{\gamma_m\}$ находятся как решения системы $(M+1)$ линейных уравнений

$$\begin{bmatrix} \hat{O}(0) & \hat{O}(\omega_1) & \hat{O}(\omega_2) & \hat{O}(\omega_3) & \dots & \hat{O}(\omega_M) \\ \hat{O}(\omega_1) & \hat{O}(0) & \hat{O}(\omega_1) & \hat{O}(\omega_2) & \dots & \hat{O}(\omega_{M-1}) \\ \hat{O}(\omega_2) & \hat{O}(\omega_1) & \hat{O}(0) & \hat{O}(\omega_1) & \dots & \hat{O}(\omega_{M-2}) \\ \vdots & \vdots & & & & \vdots \\ \hat{O}(\omega_M) & \hat{O}(\omega_{M-1}) & & & & \hat{O}(0) \end{bmatrix} \begin{bmatrix} 1 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_M \end{bmatrix} = \begin{bmatrix} \gamma_{M+1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (5.96в)$$

где $\omega_m = m\pi(2X)^{-1}$, $m = 0, 1, \dots, M$, и

$$\hat{O}(\omega) = 2\,\mathrm{Re}\{\hat{O}(\omega) \exp(-j\omega X)\}. \quad (5.96г)$$

Последнее выражение связывает промежуточные величины $\{\hat{O}(\omega_m)\}$ с выходными данными $\{O(\omega_m)\}$.

5.181. Осуждение

Изящество и эффективность решения уравнений (5.96б)—(5.96г) проявляются в его замкнутости и простоте: требуется лишь решить (5.96в) и найденные γ_m подставить в (5.96б). Благодаря этому для получения положительной оценки, согласованной с входными данными, не требуется итеративный алгоритм (в отличие от методов, описанных в разделах 5.15 и 5.16). Фактически названные простые машинные операции принципи-

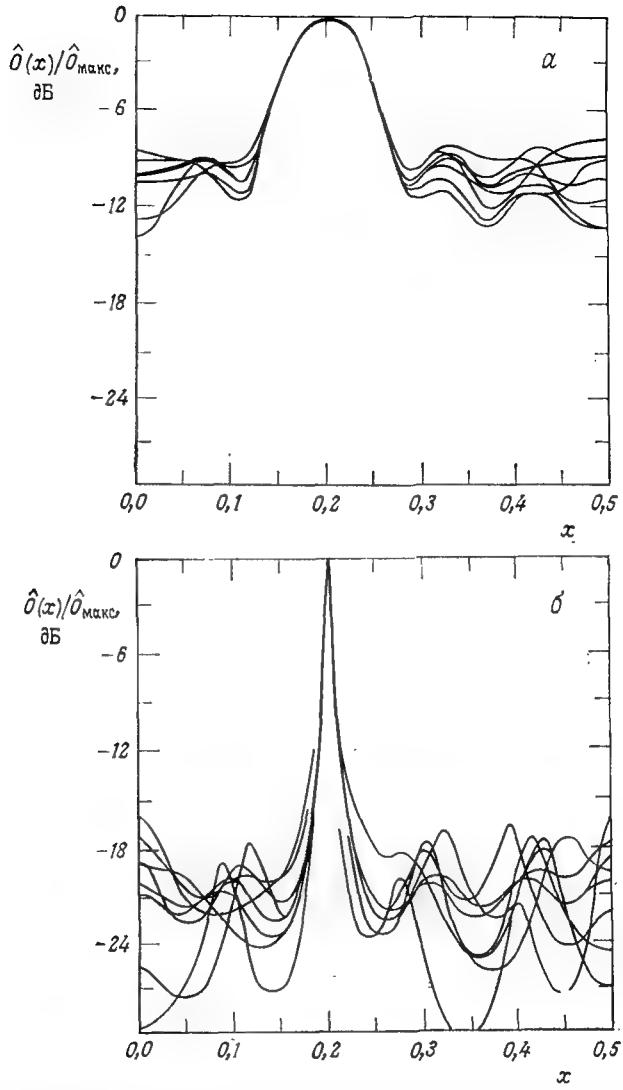
ально не отличаются от операций, которые необходимы для реализации линейных алгоритмов реставрации без ограничения по положительности; см., например, (5.26). Потенциальная экономия времени вычислений, обеспечиваемая алгоритмом Бурга, составляет его главное преимущество по сравнению с остальными алгоритмами положительной реставрации (которые являются итеративными).

Главный недостаток обсуждаемого метода состоит в его относительной чувствительности к ошибкам в данных $\hat{O}(\omega)$. Это объясняется тем, что ошибки не учитывались при выводе алгоритма. Например, в процессе вывода матрица, входящая в (5.96в), считалась неотрицательно определенной. Это было бы верно при условии, что $\{\hat{O}(\omega_m)\}$ совпадают с истинными значениями $\{O(\omega_m)\}$; однако это не обязательно выполняется при наличии ошибок во входных данных.

Эта трудность обусловила получение неоднородных результатов при фактическом использовании метода. Широкие испытания метода на зашумленных данных были проведены Лакоссом [6]. На фиг. 5.13, *б* показан пример оценки объекта, согласно (5.96б), при использовании семи наборов зашумленных данных $\hat{O}(\omega_m)$ одного и того же объекта $o(x)=1+\delta(x+0{,}2)+\delta(x-0{,}2)$. Данные были искажены равномерным случайным шумом с нулевым средним и стандартным отклонением, составляющим 3,3% от максимального сигнала. Было использовано семь разных наборов значений шума. Утолщенной линией на фигуре показан результат реставрации идеального набора данных (с нулевым шумом). Мы видим, что боковые лепестки, соответствующие ложным деталям, относительно слабы, тогда как импульсы, соответствующие пикам истинного объекта, довольно узки. Такую реставрацию можно считать достаточно хорошей, особенно по сравнению с показанной на фиг. 5.13, *а* линейной реставрацией по тем же данным с использованием треугольной функции окна (п. 5.9.2). Правда, такое сравнение несколько неравноценно.

Дело в том, что примеры реставрации по максимуму энтропии на фиг. 5.13, *б* после вычислений по алгоритму были перенормированы на единичный максимальный уровень. До перенормировки эти максимальные уровни довольно сильно отклонялись от истинного уровня, причем типичное отклонение составляло около $\pm 20\%$ (см. табл. 3 в работе [6]). Следует также заметить, что при увеличении шума от 3,3 до 5% происходило резкое увеличение ошибок.

Лакосс нашел частичное решение этой проблемы, обнаружив, что площадь центрального лепестка дает лучшую аппроксимацию истинного значения высоты пика, чем ее непосредственная оценка. Однако это верно только для очень узких импульсов



Фиг. 5.13. Сравнение реставраций, выполненных: *а* — с использованием обычной треугольной функции окна $W(\omega)$; *б* — с помощью алгоритма реставрации Бурга по максимуму энтропии (5.96б)—(5.96г). Для обоих методов представлены результаты реставрации 7 наборов зашумленных входных данных. Утолщенные кривые соответствуют случаю нулевого шума. В пределах показанной области x истинный объект имел форму дельта-функции при $x = 0{,}2$. Метод Бурга обеспечивает более высокое разрешение, чем стандартный метод. Боковые лепестки довольно малы (заметим, что используется логарифмическая шкала уровней). Однако оценки максимальных уровней существенно отличаются от истинных значений (см. текст) [6].

в объекте. Но при какой ширине импульса точность определения высоты перестает улучшаться, и как точность связана с формой импульса? И, наконец, что будет, если пользователь встретится с неизвестной структурой, содержащей импульсы и широкие плоские зоны? Вот основные вопросы, которые не дают покоя геофизикам и астрономам.

5.19. Решение Фридена для максимума энтропии

Это решение строится на основе энтропийной меры (5.95). Его дополнительное отличие от метода Бурга состоит в том, что шум на изображении рассматривается как вторая неизвестная структура, которую также следует оценить. Таким образом, в данном случае требуется еще получить несмещенную оценку шума, т. е. его оценку по максимуму энтропии.

Однако шум, присутствующий в данных об изображении, может принимать как положительные, так и отрицательные значения; при этом $\ln n_m$ для отрицательных n_m оказывается неопределенным. Эту трудность можно обойти, определив новое множество значений шума в виде

$$N(y_m) = n(y_m) + B, \quad B \geqslant 0. \tag{5.97а}$$

Здесь B — положительная постоянная достаточной величины, обеспечивающая компенсацию самых отрицательных значений $n(y_m)$ и тем самым получение всех $N(y_m) \geqslant 0$. Таким образом, мы хотим теперь найти оценки по максимуму энтропии для $N(y_m)$ и $o(x_n)$.

Вообще говоря, для всех $N(y_m)$ априорно допускаются значения от 0 до ∞. Следовательно, согласно (5.97а), в идеальном случае B должно находиться как

$$B = -\text{ наиболее отрицательное } n(y_m). \tag{5.97б}$$

Очевидно, пользователь не может точно знать это число. Однако он может принять для B некоторое разумное значение, например -2σ, если, конечно, такая априорная информация о шуме имеется в наличии. Мы убедились на опыте, что качество решения для неизвестных $\{o(x_n)\}$ и $\{N(y_m)\}$ слабо зависит от точного выполнения (5.97б), но получается все же более хорошее решение, когда это условие действительно выполняется.

Поскольку объект и шум выражаются независимыми массивами чисел, общая энтропия объекта и шума представляет собой просто сумму их энтропий. Для того чтобы решения, основанные на максимуме энтропии, отличались от идеально плоских оценок, необходимо ввести ограничения в виде входных данных об изображении и его полной энергии I_0 (см. заключительную часть раздела 5.17). В соответствии с законом

сохранения энергии величина I_0 равна общему потоку энергии от объекта. С учетом всего сказанного выше уравнение реставрации принимает вид

$$-\sum_{n=1}^{N} \hat{o}(x_n) \ln \hat{o}(x_n) - \rho \sum_{m=1}^{M} \hat{N}(y_m) \ln \hat{N}(y_m) -$$
$$-\sum_{m=1}^{M} \lambda_m \left[\sum_{n=1}^{N} \hat{o}(x_n)\, s(y_m - x_n) + \hat{N}(y_m) - B - i(y_m) \right] -$$
$$-\lambda_{M+1} \left(\sum_{n=1}^{N} \hat{o}(x_n) - I_0 \right) = \max. \qquad (5.97в)$$

Это выражение полезно сравнить с уравнением Бурга (5.96а). Новый входной параметр ρ позволяет пользователю усиливать сглаживание каждой из оценок $\{\hat{o}(x_n)\}$ и $\{\hat{N}(y_m)\}$. Чем больше ρ, тем большее значение придается максимизации энтропии шума $\{\hat{N}(y_m)\}$ и, следовательно, его сглаживанию, и наоборот. Описываемые ниже контрольные реставрации показали, что для широкого класса объектов и шума оптимум достигается приблизительно при $\rho = 20$.

Для решения (5.97в) сначала производим дифференцирование $\partial/\partial\hat{o}(x_n)$ и $\partial/\partial\hat{N}_m$ при всех m и n, а затем все результаты приравниваем нулю. Благодаря этому получаем отдельные решения для каждого из массивов неизвестных:

$$\hat{o}(x_n) = \exp\left[-1 - \lambda_{M+1} - \sum_{m=1}^{M} \lambda_m s(y_m - x_n) \right], \qquad (5.98а)$$

$$\hat{N}(y_m) = \exp(-1 - \lambda_m/\rho). \qquad (5.98б)$$

Определяющие эти решения неизвестные $\{\lambda_m\}$ находим, требуя, чтобы решения удовлетворяли входным ограничительным уравнениям в виде

$$\sum_{n=1}^{N} \hat{o}(x_n)\, s(y_m - x_n) + \hat{N}(y_n) - B = i(y_m), \quad m = 1, 2, \ldots, M, \qquad (5.98в)$$

$$\sum_{n=1}^{N} \hat{o}(x_n) = I_0.$$

Правые части (5.98в) — это входные данные. Подставив (5.98а) и (5.98б) в (5.98в), получим систему $(M+1)$ нелинейных уравнений для $(M+1)$ неизвестного $\{\lambda_m\}$.

Хотя возможны и другие пути решения, мы с успехом использовали алгоритм релаксации Ньютона — Рафсона (см., например, [55]). Вначале выбираем исходные значения $\{\lambda_m\}$; мы брали $\lambda_m = 0$ для всех $m < (M+1)$, а λ_{M+1} — таким, чтобы оно

удовлетворяло $(M+1)$-му ограничительному уравнению в системе (5.98в). Затем производим изменение $\{\lambda_m\}$ в таком направлении, которое приближает их к значениям, удовлетворяющим все ограничительные уравнения; при этом предполагается, что все производные высших порядков $\partial^{(k)} i(y_m)/\partial\lambda_p^{(k)}$ равны нулю при $k=2, \ldots$. Для получения множества $\{\lambda_m\}$, удовлетворяющего (5.98в), требовалось от 8 до 40 итераций.

5.19.1. Обсуждение

Интересно отметить, что, хотя мы начинали вывод с априорного предположения о максимуме энтропии (5.97в), в применении к данной задаче это исходное положение само может быть выведено на основе принципа максимального правдоподобия [18]. Поэтому решения (5.98а) и (5.98в) одновременно являются оценками по максимуму энтропии и по максимуму правдоподобия.

Что же касается положительности, то, как видно из (5.98а), сама форма представления решения гарантирует положительную оценку объекта.

В отношении паразитных осцилляций на выходе $\hat{o}(x)$ из (5.98а) следует, что

$$d^{(k)}\hat{o}(x)/dx^{(k)} \infty \hat{o}(x), \quad k=1, 2, \ldots .$$

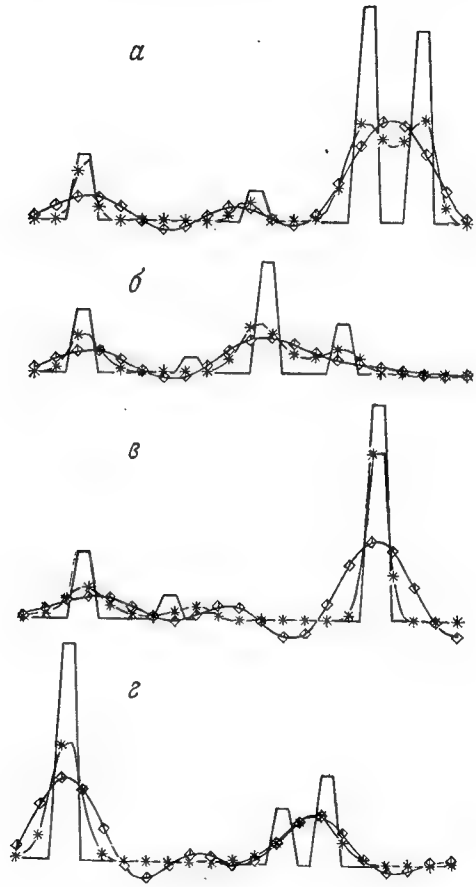
Это означает, что в тех местах, где имеется нулевой фон, $\hat{o}(x)$ вообще не может осциллировать либо осциллирует слабо. Практически в пределах тех областей, где объект $\hat{o}(x)$ был равен нулю, типичные оценки $o(x)$ везде давали значения порядка 10^{-8}. Поэтому в таких областях оценки были очень гладкими, а осцилляции совершенно отсутствовали. Именно по этой причине метод дает вполне хорошие результаты для объектов, содержащих множество импульсов и имеющих нулевые значения везде, кроме конечного числа точек.

5.19.2. Контрольные реставрации

Фиг. 5.14 иллюстрирует применение описанного метода для реставрации импульсной структуры, размытой под действием дифракции. Представленные данные были получены путем моделирования на ЭВМ.

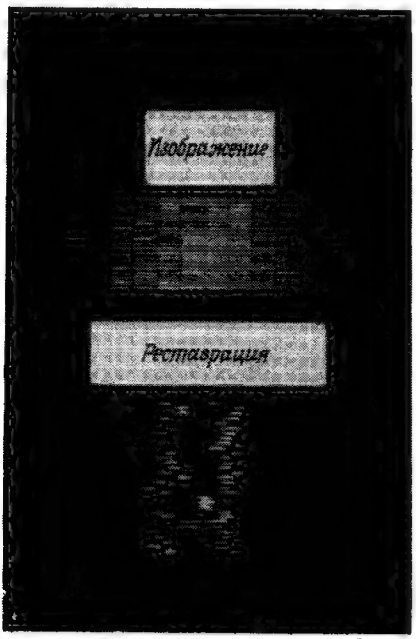
На фиг. 5.15 представлен результат реставрации фотографии двойной щели, размытой из-за дифракции. Щели находились на расстоянии, равном половине релеевского расстояния разрешения, что исключало реставрацию линейными методами. Как видно из фотографии, изображение было искажено довольно сильным шумом; оценочное значение стандартного отклонения шума

составляло $\sigma \approx 21\%$. Примечательно, что действие шума на каждую строку реставрации проявилось лишь в пространственном смещении пары импульсов, которые были восстановлены почти на всех строках. Из-за этого реставрация в целом создает зрительное впечатление колеблющейся двойной щели.

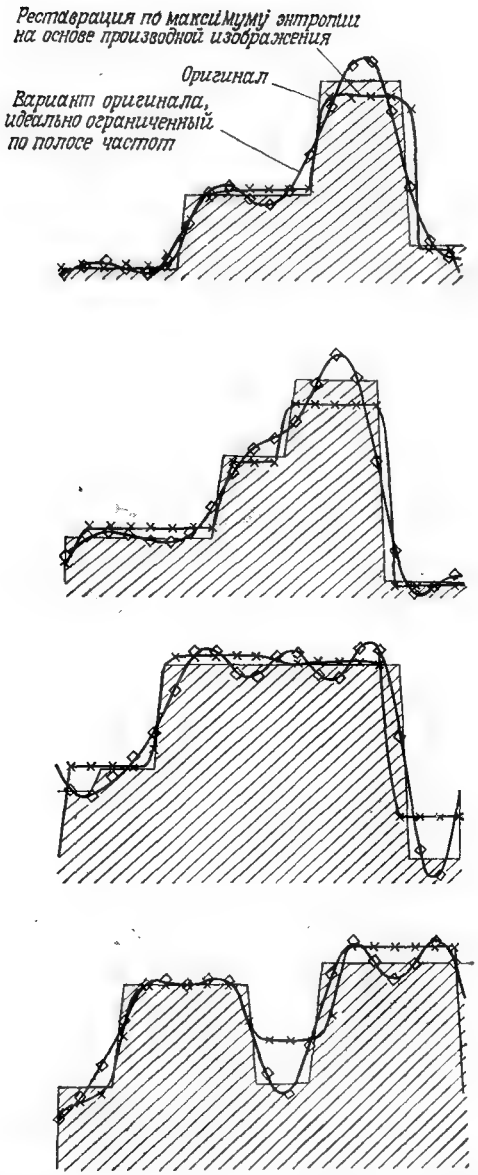


Фиг. 5.14. Реставрации по методу максимума энтропии Фридена (кривые со звездочками) некоторых импульсных объектов (сплошные кривые). Использовано значение $\rho = 1$. Не показанные на фигуре входные данные $\{i(y_m)\}$ генерировались с помощью ЦВМ и искажались чисто дифракционным смазыванием и равномерным случайным шумом с максимальной относительной ошибкой 5%. Чтобы продемонстрировать преимущества данного метода по сравнению с линейными методами реставрации, показаны также реставрации, полученные с помощью инверсной фильтрации в отсутствие шума (кривые с квадратиками).

В применении к изображениям объектов, состоящих из случайных ступенек, рассматриваемый метод дает лишь несколько более хороший результат, чем метод оптимальной линейной фильтрации. Главным недостатком метода в этом случае является появление таких же осцилляций из-за явления Гиббса, какие присущи линейной реставрации. Однако удалось найти способ преодоления этой трудности. Если объект состоит из случайных ступенек, то его производная содержит случайные импульсы, которые хорошо реставрируются описанным методом (см. выше). Поэтому, когда пользователь знает, что интересующий его объект состоит из случайных ступенек, он может продифференцировать данные об изображении и ввести полученные



Фиг. 5.15. Метод максимума энтропии Фридена в приложении к сильно зашумленным данным об изображении. Шум характеризуется величиной $\sigma \approx \approx 0{,}2i_{\text{макс}}$. Объект, состоящий из двух щелей, разнесенных на $^1/_2$ релеевского расстояния разрешения, был размыт в направлении поперек щелей в результате чистой дифракции, создаваемой оптической апертурой щели. Полученное изображение фотографировалось, а затем сканировалось и реставрировалось построчно. Было использовано значение $\rho = 20$. Показаны первые 30 строк входных данных и реставрации. Хотя почти на каждой строке происходит желаемая реставрация двух импульсов, сильный шум в данных приводит к возникновению случайных смещений импульсов. Однако эти смещения невелики — порядка 1/10 релеевского расстояния разрешения [63].



Фиг. 5.16. Применение метода максимума энтропии Фридена к объектам в виде случайных ступенек. Данные об изображении, полученные с помощью ЦВМ, размывались с использованием ядра вида $\mathrm{sinc}^2(x)$ и затем дискретизировались с шагом, равным $^1/_2$ интервала Найквиста. Такая более высокочастотная, чем обычно, дискретизация обеспечивала получение довольно хорошей аппроксимации первой производной конечными разностями. Метод пригоден для использования в тех случаях, когда пользователь заранее знает, что объект состоит из случайных ступенек.

числа в алгоритм. При этом выход будет иметь форму $d\hat{o}/dx$. Восстановление $\hat{o}(x)$ производится путем интегрирования по полю объекта.

Мы произвели проверку метода для подобных объектов, результаты которой представлены на фиг. 5.16. Изображения (не показанные на фигуре) дифференцировались с помощью конечных разностей между соседними значениями, отстоящими на половину интервала Найквиста. Сравнение с реставрацией тех же объектов при идеальном ограничении полосы частот в соответствии с (5.8) демонстрирует преимущества данного метода: а) гораздо более крутые переходы на границах и б) полное устранение паразитных осцилляций. В отношении способности идентифицировать наличие границ и площадок без появления ложных деталей этот алгоритм также дает лучший результат, чем оценка при идеальном ограничении полосы частот. Однако это оказывается неверным, если производить сравнение на основе среднеквадратичной ошибки. Дело в том, что по методу максимума энтропии высоты площадок определяются обычно неточно. Это объясняется накоплением ошибок при интегрировании по полю объекта, содержащего реставрированные дельта-функции, площади которых подвержены случайным ошибкам. Отсюда можно сделать вывод, что предложенный метод следует использовать в тех случаях, когда более важно определить геометрическую конфигурацию объекта, чем абсолютные уровни яркости.

5.20. Метод реставрации Монте-Карло

Все методы реставрации, описанные выше в этой главе, основаны на использовании функционального выражения оценки $\hat{o}(x)$, которая в конечном счете каким-то образом связана с входами. Эти функциональные выражения содержат решения либо в явной форме, как (5.61), либо в неявной форме, как (5.88).

Возможно, что причина преобладания таких функциональных методов состоит в том, что человеку свойственна природная склонность к аналитическому решению задач. Статистические методы весьма часто рассматриваются как последнее средство, к которому обращаются лишь после неудачи с попытками применить аналитические методы. Еще одна причина состоит просто в том, что неумение пользоваться статистическими методами предопределяется пробелами в образовании многих специалистов в данной отрасли (исключая инженеров-электриков, но включая физиков!).

Одним из широко используемых инструментов статистики является вычислительный метод Монте-Карло. Он основан на том, что представительный образец (например, объект) может

быть построен на ЦВМ с помощью большого количества испытаний, производимых случайно, но в соответствии с вероятностными и детерминистическими законами, определяющими возможные выходы (т. е. объекты).

Чтобы приложить этот принцип к решению задач реставрации, вначале нужно представить, что объектная область составлена из «пустых» клеток разрешаемого размера. Затем с помощью определенного правила принятия решения по этим клеткам случайно распределяются математические «зерна» с фиксированным приращением интенсивности do. При этом реставрируемый объект строится путем последовательных распределений зерен, пока все они не будут размещены по клеткам. (Общее число зерен может быть определено из данных об изображении на основе сохранения энергии.)

Такой подход имеет ряд потенциальных преимуществ. Во-первых, освободившись от аналитического выражения, выход $\hat{o}(x)$ приобретает неограниченную приспосабливаемость к флуктуациям; крутизна переходов и линейное расстояние разрешения теперь уже не обусловливаются полосой частот или другими факторами. В некотором смысле такой подход является предельно нелинейным.

Во-вторых, все желаемые ограничения (такие, как требование положительности, верхний предел и т. п.) здесь могут быть легко учтены просто путем исключения определенных событий из множества возможных. Насколько важную роль играет наложение подобных ограничений, мы уже видели.

В-третьих, метод в принципе обеспечивает очень высокую скорость вычислений; благодаря этому, наконец, появляется возможность реставрировать с повышенным разрешением двумерные объекты, тогда как ранее это удавалось только для одномерных объектов. Скорость вычисления в конечном счете определяется используемым правилом принятия решения, от которого зависит попадание зерна в данную клетку.

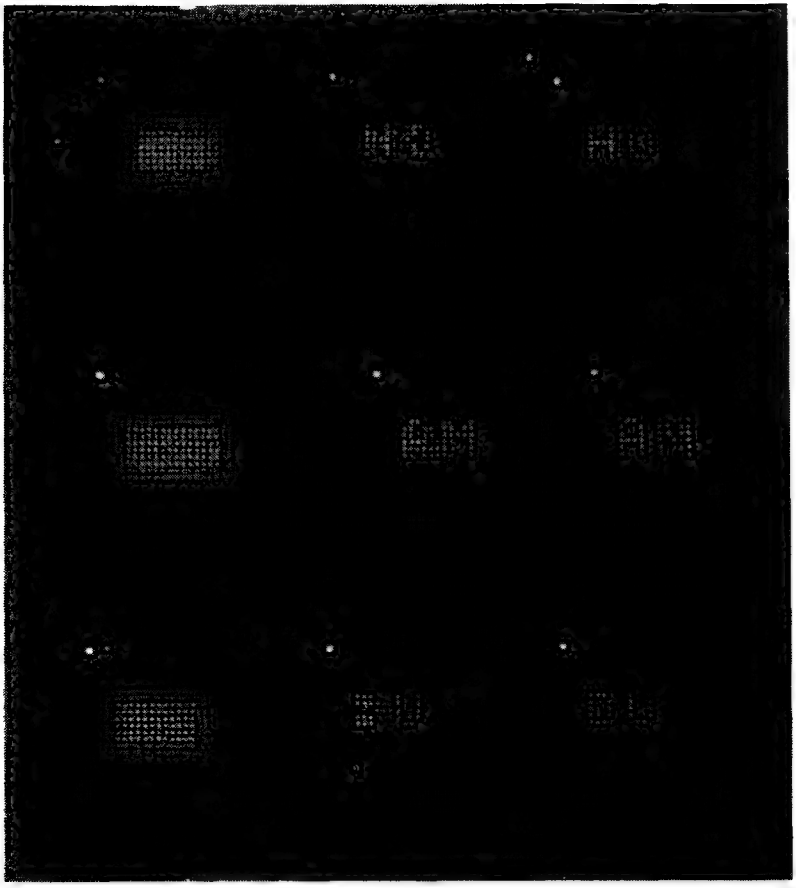
Рассмотрим правило принятия решения, которое, по-видимому, пригодно для поставленной задачи [19]. Пусть зерно k попадает в клетку x_n, если получаемое в результате накопления изображение $\hat{i}^{(k)}(y_m)$, которое формируется в соответствии с

$$\hat{i}^{(k)}(y_m) \equiv \hat{i}^{(k-1)}(y_m) + do \cdot s(y_m - x_n), \tag{5.99а}$$

удовлетворяет условию

$$\hat{i}^{(k)}(y_m) \leqslant r \cdot i(y_m) \quad \text{при всех} \quad m = 1, \ldots, M \tag{5.99б}$$

с минимальным r. Минимальное r означает, что при еще меньшем r зерно k в соответствии с (5.99а) уже не сможет попасть ни в одну из клеток x_n. Как и выше, здесь $\{y_m\}$ — координаты



Фиг. 5.17. Образец реставрации с использованием распределения «зерен» по методу Монте-Карло. В каждом горизонтальном ряду показаны данные об изображении, реставрация и истинный объект. Данные об изображении не содержали шума. Детали букв нельзя было восстановить с помощью линейной фильтрации. Как следует из прямого сравнения, метод Монте-Карло обеспечивает или почти обеспечивает распознавание букв [19].

данных, а $\{i(y_m)\}$ — данные об изображении. Когда зерно k попадает в клетку x_p, объект увеличивается до значения

$$o^{(k)}(x_n) = o^{(k-1)}(x_n) + do \cdot \delta_{np}. \qquad (5.99в)$$

Алгоритм (5.99а)—(5.99в) весьма эффективен для расчетов на ЭВМ. Он состоит из многократных пробных операций суммирования (5.99а) и сравнения (5.99б), после которых производится

еще одно суммирование (5.99в), когда зерно попадает в клетку. Заметим, что дискретизированную функцию рассеяния точки $s(y_m - x_n)$, входящую в этот алгоритм, можно вычислить и запомнить на начальном этапе программы, с тем чтобы при последующем применении алгоритма извлекать эту функцию из памяти в готовом виде.

Фиг. 5.17 иллюстрирует применение метода для реставрации сильно смазанного печатного текста. Функция рассеяния точки имела вид $\mathrm{sinc}^2(x) \cdot \mathrm{sinc}^2(y)$. Смазывание приводило к тому, что ширина каждой буквы и пробела между буквами оказались равными половине релеевского расстояния разрешения. При таком смазывании инверсная фильтрация не позволяет разрешить буквы даже в отсутствие шума в данных об изображении. Это означает, что имеющейся полосы частот недостаточно и требуется экстраполяция.

Каждое входное изображение представляло собой матрицу 21×21 с шагом, равным $^3/_4$ интервала Найквиста. Поскольку объекты были двоичными с уровнями 0 и 1, мы использовали значение $do = 0{,}2$. Таким образом, для заполнения клетки требовалось 5 зерен. В показанных примерах применения алгоритма шум в данные не вводился.

Затем были проведены вычисления с теми же данными, но при $do = 0{,}1$. Хотя время, требуемое для реставрации пары букв, увеличилось приблизительно от 1 мин до 2 мин, точность реставрации заметно возросла. При введении шума, характеризуемого равномерной случайной ошибкой с максимальным значением 5% от уровня входного изображения, качество реставрации снижалось и становилось примерно таким, как показано на фиг. 5.17.

Разработка этого алгоритма еще не окончена.

5.21. Прочие методы

За последние годы в открытой литературе появились описания множества предложенных методов реставрации. В предыдущих разделах были рассмотрены те методы, относительно которых у автора имелись достаточные теоретические или экспериментальные сведения, чтобы сделать изложение понятным и полезным.

Ряд методов, которые нам пока еще не представилось случая исследовать, приводятся в работах Ричардсона [56], Мак-Адама [57] и Савчука [58]. Более тщательный обзор позволит, очевидно, обнаружить и другие методы.

До сих пор мы всегда полагали, что ФРТ $s(x)$ является детерминированной функцией, известной с некоторой точностью.

Однако в случае атмосферной турбулентности более правильно рассматривать $s(x)$ как случайную переменную. Для этого случая Слепян [59] определил реставрирующий фильтр на основе МСКО, который представляет собой обобщение фильтра (5.68), полученного для детерминированной $s(x)$.

Большинство обрабатываемых изображений имеет форму фотографий, а большинство реставраций предназначается для последующего визуального изучения. Понятно, что при любой реставрации фотографий полезно уменьшать зернистость и оптимизировать изображение в расчете на зрительное восприятие. Однако все эти проблемы выходят за рамки данной главы. Соответствующие сведения можно получить, например, из работ Хуанга [60] и Стокхэма [61].

5.22. Сводка результатов

Со студенческих дней мы всегда ждем, что в конце все результаты будут сжаты до одной-двух страниц. По существу дела такое сжатое представление позволяет убрать те самые деревья, из-за которых не видно леса. В табл. 5.1 приводится сводка всех рассмотренных методов. Порядок перечисления такой же, как в основном тексте. Для каждого метода реставрации указывается, к какому типу он принадлежит (линейному или нелинейному, с ограниченной или неограниченной полосой частот, с учетом или без учета физических ограничений), приводятся данные по разрешению и указываются возможности применения для реставрации больших двумерных изображений. Такое взаимное сопоставление методов учитывает интересы пользователей. Наконец, в последнем столбце даются различные примечания.

5.23. Возможные направления будущих исследований

Анализ таблицы показывает, что разрешение порядка $R/2$ достигается довольно легко. Как объясняется в п. 5.13.2, это значение приблизительно соответствует практическому пределу всех линейных методов реставрации без учета ограничений. Только новые, нелинейные методы позволяют заметно уменьшить предельное расстояние разрешения — практически приблизительно до $R/4$.

С другой стороны, главной областью применения нелинейных методов является реставрация таких импульсных объектов, как звезды, линейчатые спектры и т. п. Эти способы оказываются не

Таблица 5.1

Основные характеристики методов реставрации

Метод (название или автор)	Тип [1])	Раздел или пункт	Разрешение (x_0 — расстояние разрешения, R — релеевское расстояние)	Обработка больших двумерных изображений (матриц $M \times M$)	Общие замечания
Инверсия матричного уравнения отображения	1 НП НИ	5.3.1	Теоретически идеальное, практически $x_0 \geqslant R/2$	Требует решения матрицы $M^2 \times M^2$	Сверхчувствителен к шуму в данных
Филлипс	То же	5.4	Теоретически идеальное, практически компромисс разрешения и ложных деталей	То же	Вводится параметр сглаживания γ, контролирующий вторые разности
Туми	» »	5.5	То же	То же. Входная матрица может быть прямоугольной	Общий метод квадратичного сглаживания
Инверсная фильтрация	1 П НИ	5.9	$x_0 = R/2$	Применение продемонстрировано	В сочетании с алгоритмом БПФ обеспечивает высокую скорость
То же с оптимальной полосой частот $2\Omega_p$	То же	5.9.1	В частном случае $x_0 = (R/2)(1 - \phi_n/\phi_0)^{-1/2}$	Трудностей не ожидается	В общем случае для Ω_p должно выполняться $\lvert \tau(\Omega_p) \rvert^2 = \phi_n(\Omega_p)/\phi_0(\Omega_p)$
Фильтрация с применением функции окна	» »	5.9.2	Практически $x_0 = R/2$	То же	Обеспечивает компромисс разрешения и ложных деталей
Подавление постоянной составляющей в полосе частот $2\omega_0$	» »	5.10.1	$x_0 \to R/4$ при $\omega_0 \to \Omega$ (Ω — оптическая граничная частота)	» »	Высокий уровень ложных деталей при разрешении $x_0 \leqslant R/2$
Чистая фазовая компенсация	» »	5.10.3	$x_0 \approx R$	» »	Относительно нечувствителен к шуму в данных
ван Циттерт	1 П И	5.10.4	$x_0 = R/2$	» »	При конечном числе итераций k приближается к фильтрации с функцией окна

Дискретная свертка	1 П НИ	5.10.5	Практически $x_0 \geqslant R/2$	Применение продемонстрировано	Пространственный эквивалент фильтрации в частотной области
Винер — Хелстром	То же	5.11.1	$x_0 \geqslant R/2$	Трудностей не ожидается	«Стандартный» метод МСКО-фильтрации
То же с ограничением резкости	» »	5.11.1	То же	То же	Позволяет ограничить резкость выхода
Бакус — Гильберт	1 НП	5.11.2	В принципе обеспечивает гибкое управление разрешением, практически $x_0 \geqslant R/2$	Требует решения матрицы $M^2 \times M^2$	Обеспечивает оптимальный компромисс усиления шума и разрешения
Фильтровый вариант метода Бакуса — Гильберта	1 П НИ	5.11.2	То же	Трудностей не ожидается	То же
Линейные методы экстраполяции (Харрис, Барнс, Фриден)	1 НП НИ	5.13.1 5.13.2	Теоретически идеальное, но практически $x_0 \geqslant R/2$	Возможна, но, вероятно, не оправдает усилий	Особые трудности из-за излишней чувствительности к шуму
Шелл и Биро	2 НП И	5.15	Практически достигнуто $x_0 \approx R/4$	Возможна	Обеспечивает наилучшие результаты для импульсных объектов
Джанссон — ван Цитгерт	То же	5.16	То же	Вполне возможна	То же
Бург	2 НП НИ	5.18	» »	Требует решения тёплицевой матрицы $M^2 \times M^2$	» »
Фриден (максимум энтропии)	2 НП И	5.19	» »	Требует итеративного решения матрицы $M^2 \times M^2$	» »
Фриден (Монте-Карло)	То же	5.20	» »	Возможность продемонстрирована	» »

*) Обозначения: 1 — выход линейно связан с входными данными; ограничения отсутствуют;
2 — выход нелинейно связан с входными данными; используется ограничение по положительности;
П — ограниченная полоса частот;
НП — неограниченная полоса частот;
И — итеративный алгоритм;
НИ — неитеративный алгоритм.

слишком эффективными для объектов, содержащих случайные ступеньки или переходы. Для подобных объектов довольно хорошие результаты получаются с помощью дифференцирования данных об изображении (п. 5.19.2), однако этот прием применять нежелательно, поскольку он приводит к росту ошибок реставрации из-за усиления шума, а также поскольку пользователь должен знать заранее, что имеет дело с объектом ступенчатого типа. Наконец, все нелинейные методы (исключая метод Бурга) имеют еще один недостаток — они реализуются итеративными алгоритмами.

Поэтому главные усилия должны быть, видимо, направлены на создание алгоритмов, способных хорошо реставрировать как импульсные, так и ступенчатые объекты. Желательно, чтобы эти алгоритмы обеспечивали получение решений в замкнутой, а не в итеративной форме. Может быть, в последующие годы будет достигнут прогресс в решении этих проблем. Однако, как абсолютно ясно уже теперь, для того, чтобы идти в ногу с таким прогрессом, потребуется отказаться от узко профессионального подхода и следить за научной литературой по многим направлениям, которые на первый взгляд кажутся не относящимися к делу.

Литература

1. *Phillips D. L.*, J. ACM, **9**, 84 (1962).
2. *Twomey S.*, J. Franklin Inst., **179**, 95 (1965).
3. *Yamamoto G., Tanaka M.*, Appl. Opt., **8**, 447 (1969).
4. *Vanasse G. A., Sakai H.*, Progress in Optics VI, ed. by E. Wolf, North-Holland Publishing Co., Amsterdam, 1967.
5. *Burg J. P.*, Maximum Entropy Spectral Analysis (Stanford University, Geophysics Department) (paper presented at 37th Annual Society of Exploration Geophysicists Meeting, Oklahoma City, 1967).
6. *Lacoss R. T.*, Geophysics, **36**, 661 (1971).
7. *Gordon R., Herman G. T.*, Comm. ACM-Graphics and Image Processing, **14**, 759 (1971).
8. *Schell A. C.*, Radio Electronic Eng., **29**, 21 (1965).
9. *Biraud Y.*, Astron. Astrophys., **1**, 124 (1969).
10. *Jansson P. A., Hunt R. H., Plyler E. K.*, J. Opt. Soc. Am., **60**, 596 (1970).
11. *Jaynes E. T.*, IEEE Trans. Syst. Sci. Cybern., SSC-4, 227 (1968).
12. *Wragg A., Dowson D. C.*, IEEE Trans. Inf. Theory, IT-**16**, 226 (1970).
13. *O'Neill E. L.*, Introduction to Statistical Optics (Addison-Wesley Publishing Company, Reading, Massachusetts, 1963). (Имеется перевод: О'Нейл Э. Введение в статистическую оптику. — М.: Мир, 1966.)
14. *Twomey S.*, J. ACM, **10**, 97 (1963).
15. *Strand O. N., Westwater E. R.*, SIAM J. Numer. Anal., **5**, 287 (1968).
16. *Pratt W. K.*, Ed., Bibliography on Digital Image Processing and Related Topics, Univ. of Southern California-USCEE Report 453, Los Angeles, 1973.
17. *Huang T. S., Schreiber W. F., Tretiak O. J.*, Proc. IEEE, **59**, 1568 (1971). [Имеется перевод: ТИИЭР, **59**, № 11, 59 (1971).]
18. *Frieden B. R.*, J. Opt. Soc. Am., **62**, 511 (1972).

19. *Frieden B. R.*, Image Restoration by Decision-Rule Allocations of Pseudo-Grains, paper presented at Spring Meeting of the Optical Society of America, Washington, D. C., 1974.
20. *Good I. J.*, J. Roy. Stat. Soc., **B20**, 361 (1958); **B22**, 372 (1960).
21. *Cooley J. W., Tukey J. W.*, Math. Computation, **19**, 297 (1965).
22. *Singleton R. C.*, IEEE Trans. Audio Electroacoustics, **AU-15**, 91 (1967).
23. *McGlamery B. J.*, J. Opt. Soc. Am., **57**, 293 (1967).
24. *Papoulis A.*, Probability, Random Variables, and Stochastic Processes, McGraw-Hill Book Co., New York, 1965.
25. *Davenport W. B., Root W. L.*, Random Signals and Noise, McGraw-Hill Book Co., New York, 1958. (Имеется перевод: *Давенпорт В. Б., Рут В. Л.* Введение в теорию случайных сигналов и шумов. — М.: ИЛ, 1960.)
26. *Sondhi M. M.*, Proc. IEEE, **60**, 842 (1972). [Имеется перевод: ТИИЭР, **60**, № 17, 108 (1972).]
27. *Harris J. L.*, J. Opt. Soc. Am., **56**, 569 (1966).
28. *Arguello R. J., Sellner H. R., Stuller J. A.*, IEEE Trans. Computers, **C-21**, 812 (1972).
29. *Jacquinot P., Roizen-Dossier B.*, в книге Progress in Optics III, ed. by E. Wolf, North-Holland Publishing Company, Amsterdam, 1964.
30. *Frieden B. R.*, Optica Acta, **16**, 795 (1969).
31. *O'Neill E. L.*, IRE Trans. Inf. Theory, **2**, 56 (1956).
32. *Swindell W.*, Appl. Opt., **9**, 2459 (1970).
33. *Hawman E. W.*, Internal report, Optical Sciences Center, University of Arizona, 1973.
34. *Baker R., Burke J., Frieden R.*, Progress in Digital Image Processing, 1969 (Techn. Report 50, Optical Science Center, University of Arizona, 1970).
35. *Van Cittert P. H.*, Z. Phys., **69**, 298 (1931).
36. *Jansson P. A., Hunt R. H., Plyler E. K.*, J. Opt. Soc. Am., **58**, 1665 (1968).
37. *Frieden B. R.*, J. Opt. Soc. Am., **64**, 682 (1974).
38. *Nathan R.*, In Pictorial Pattern Recognition (G. C. Cheng, Thompson, Washington, D. C., 1968).
39. *Helstrom C. W.*, J. Opt. Soc. Am., **57**, 297 (1967).
40. *Fellgett P. B., Linfoot E. H.*, Phil. Trans. Roy. Soc. (London), **A 247**, 369 (1955).
41. *Frieden B. R.*, J. Opt. Soc. Am., **60**, 575 (1970).
42. *Backus G., Gilbert F.*, Phil. Trans. Roy. Soc. (London), **A 266**, 123 (1970).
43. *Saleh B. E. A.*, Appl. Opt., **13**, 1833 (1974).
44. *Wolter H.*, в книге Progress in Optics I, ed. by E. Wolf, North-Holland Publishing Co., Amsterdam, 1961.
45. *Harris J. L.*, J. Opt. Soc. Am. **54**, 931 (1964).
46. *Frieden B. R.*, Appl. Opt., **9**, 2489 (1970).
47. *Slepian D., Pollak H. O.*, Bell Syst. Tech. J., **40**, 43 (1961).
48. *Frieden B. R.*, In Progress in Optics IX, ed. by E. Wolf, North-Holland Publishing Co., Amsterdam, 1971.
49. *Barnes C. W.*, J. Opt. Soc. Am., **56**, 575 (1966).
50. *Rushforth C. K., Harris R. W.*, J. Opt. Soc. Am., **58**, 539 (1968).
51. *Buck G. J., Gustincic J. J.*, IEEE Trans. Antennas Propagation, **AP-15**, 376 (1967).
52. *Hildebrand F. B.*, Introduction to Numerical Analysis, McGraw-Hill Book Co., New York, 1956, p. 443.
53. *Wong H. F.*, A Study and Evaluation of Biraud's Deconvolution Algorithm (Masters Thesis, Electrical Engineering Dept., Queen's University, Kingston, Ontario, Canada, 1971).
54. *Edward J. A., Fitelson M. M.*, IEEE Trans. Inf. Theory, **IT-19**, 232 (1973).
55. *Pipes L. A.*, Mathematics for Engineers and Physicsts, McGraw-Hill, Book Co., New York, 1958, p. 116.
56. *Richardson W. H.*, J. Opt. Soc. Am., **62**, 55 (1972).

57. *MacAdam D. P.*, J. Opt. Soc. Am., **60**, 1617 (1970).
58. *Sawchuk A. A.*, Proc. IEEE, **60**, 854 (1972). [Имеется перевод: ТИИЭР, **60**, № 7, 124 (1972).]
59. *Slepian D.*, J. Opt. Soc. Am., **57**, 918 (1967).
60. *Huang T. S.*, In Woods Hole Summer Study on Restoration of Atmospherically Degraded Images, Vol. 2 (Defense Documentation Center, Alexandria, Virginia, 1966).
61. *Stockham T. G.*, Proc. IEEE, **60**, 828 (1972). [Имеется перевод: ТИИЭР, **60**, № 7, 93 (1972).]
62. *O'Handley D. A., Green W. B.*, Proc. IEEE, **60**, 821 (1972). [Имеется перевод: ТИИЭР, **60**, № 7, 84 (1972).]
63. *Frieden B. R.*, J. Opt. Soc. Am., **62**, 1202 (1972).

6. Влияние шума аппаратуры цифровой обработки изображений

Ф. Биллингсли

В системах обработки изображений ограничивающим фактором является шум. Поэтому при проектировании любой системы формирования изображения необходимо тщательно рассматривать ее шумовые характеристики. Особое внимание следует обратить на следующие два аспекта: каким образом шум изменяет вид изображения и в каком отношении он затрудняет последующую цифровую обработку.

Ниже кратко рассматриваются некоторые области, в которых эти аспекты имеют важное значение. Однако в связи с тем, что в данной главе основное внимание уделяется рассмотрению источников шума как таковых, а не изучению конкретных проявлений шума в последующей обработке, приведенные примеры не содержат технических подробностей.

Влияние шума на качество реставрации и улучшения изображений цифровым способом оценивается по статистике возмущений, вызываемых шумом. Поэтому анализируется влияние шума на точность квантования, а полученные кривые используются для оценивания качества системы по вероятности присвоения правильного двоичного числа (ДЧ) в процессе аналого-цифрового преобразования.

Полученные данные позволили предложить концепцию некоторого инвариантного параметра (β — отношение шага квантования к среднеквадратичному значению шума), который может использоваться при оценивании шумовых характеристик системы. В работе анализируется сканер бегущего луча, в котором фотодетектором служит фотоэлектронный умножитель. Анализ ряда «нормальных» систем позволил выявить их сильные и слабые стороны, а также предложить новую систему сканирования, которая обладает лучшими свойствами каждой из прежних систем.

6.1. Влияние шума на зрительное восприятие и анализ изображения

Поскольку эти два аспекта рассмотрения (зрительное восприятие и анализ изображения) приводят к несколько различным требованиям в отношении шума, их следует разделить.

Приведенные здесь примеры не претендуют на полноту освещения проблемы, однако они создают основу для обсуждений. Автор надеется, что эти примеры позволят читателям исследовать и другие области, в которых может оказаться полезным прогнозирование шумовых характеристик.

6.1.1. Зрительное восприятие изображения

Влияние двумерного шума на зрительное восприятие наблюдателя изучалось очень многими исследователями (см., например, [1—6]). Визуальные проявления шума тесным образом связаны с многими атрибутами изображения, такими, как требуемая разрешающая способность, характер, контраст, окружающий фон и яркость, а также с его собственными характеристиками — уровнем, спектральным распределением и др. Сугубо ориентировочное эмпирическое правило (следует предупредить о возможности недопустимого упрощения) гласит, что отношение сигнал/шум 40 дБ (100 : 1) обеспечивает получение сравнительно «чистого» изображения. Более интенсивный шум понижает способность наблюдателя различать участки разной яркости (или разной прозрачности в случае фотопленки) [7, 8].

При определенных условиях шум может дать и полезный эффект. Робертс [9], например, осуществил сжатие данных благодаря разрушению раздражающих ложных контуров, возникающих в случае малоуровневого квантования, путем добавления и вычитания идентичного псевдослучайного шума на передающей и приемной сторонах. Однако этот метод, основанный на использовании физиологических особенностей наблюдателя, не принадлежит к методам, обеспечивающим сохранение информации; его применение может неблагоприятно сказаться на возможной последующей обработке.

Шум накладывает существенное ограничение на обработку изображений, наблюдаемых человеком. Рассмотрим такой пример. Изображения, принятые со спутника ERTS, могут обладать очень малым контрастом, поэтому требуется усиливать контраст в пять и более раз. При таком большом повышении контраста визуально «чистое» изображение становится крайне зашумленным. Соответственно было установлено, что увеличение уровня высших пространственных составляющих с целью повышения визуальной резкости изображения ограничено небольшим коэффициентом, например 5 (другими словами, амплитуда высших пространственных частот увеличивается в 5 раз относительно амплитуды низких частот); при большем увеличении уровня шум становится недопустимо заметным. Это накладывает ограничение на повышение резкости изображения, которое можно практически реализовать.

Конечно, визуальное проявление шума имеет значение, однако с точки зрения рассмотрения данных, участвующих в обработке изображения, этот фактор следует считать второстепенным. Учитывая это, визуальный эффект шума подробно не рассматривается.

6.1.2. Влияние шума на обработку данных

Двумерный шум изображения затрудняет измерение модуляционной передаточной функции (МПФ) [10—13]. Если производятся визуальные измерения, то имеет место понижение способности наблюдателя опознавать известные ему структуры (миры с прямоугольными или синусоидальными штрихами) с малыми размерами и низким контрастом, когда они сильно зашумлены. Переход на инструментальные измерения при помощи микроденситометра с узкой щелью позволяет уменьшить уровень регистрируемого шума путем усреднения яркостей элементов изображения по длине штрихов. Проанализировать результат сканирования щелью (способом оценивания или прибегнув к машинному анализу) можно путем выделения основной гармоники каждой пространственной структуры. Машинный метод выделения основной гармонической составляющей из зашумленного сигнала обладает достаточно высокой эффективностью, позволяющей использовать сканирующее пятно вместо сканирующей щели большой длины (хотя даже в последнем случае усреднение нескольких строк обеспечивает более «чистый» результат).

Блэкман показал искажающее действие шума на измеряемую кривую МПФ [14]. Если обработка изображения основана на знании формы этой кривой (как в некоторых случаях увеличения уровня высоких частот изображения), то искажение кривой приводит к ошибкам в последующей обработке.

Поскольку энергетический спектр шума наложен на энергетический спектр изображения, шум и полезный сигнал невозможно разделить. Поэтому обработка в спектральной области с целью реставрации спектра изображения приводит к усилению шума. Этот эффект можно минимизировать путем соответствующей модификации формы корректирующей характеристики, однако для этого необходимо знать форму энергетических спектров шума и изображения. Пусть ПФ(И) и ПФ(О) — двумерные преобразования Фурье изображения и объекта; ОПФ — оптическая передаточная функция (комплексная функция, модуль которой есть МПФ). При отсутствии шума формирование изображения можно представить в виде

$$\text{ПФ(И)} = \text{ОПФ} \cdot \text{ПФ(О)}.$$

Если ОПФ существует на всех полезных пространственных частотах, неискаженное восстановление объекта по изображению

можно осуществить согласно соотношению

$$\text{ПФ}(\text{О}) = \text{ПФ}(\text{И})/\text{ОПФ}.$$

При наличии шума изображение модифицируется «несвернутым» сигналом, поэтому реставрация приводит к свертке с шумом и появлению шума в реставрированном сигнале. Реставрацию при наличии шума можно оптимизировать путем винеровской фильтрации, причем соответствующая реставрирующая функция имеет вид [15]

$$\text{ПФ}(\text{О}) = [\text{ПФ}(\text{И})/\text{ОПФ}]\text{ЭС}(\text{С})/[\text{ЭС}(\text{С}) + \text{ЭС}(\text{Ш})],$$

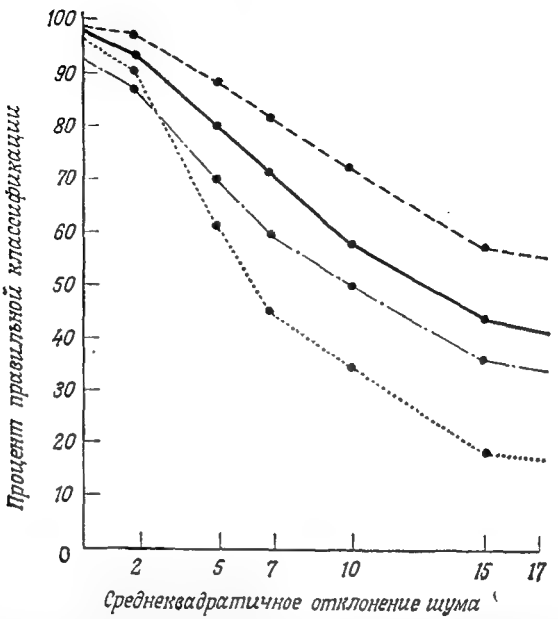
где ЭС(С) и ЭС(Ш) — энергетические спектры сигнала и шума (предполагается, что они некоррелированы). Обычно ЭС(С) и ЭС(Ш) неизвестны, поэтому наличие шума вносит трудность в том отношении, что требуется определить или оценить эти энергетические спектры, как правило, зная только ЭС(С + Ш).

Кодированию и влиянию шума посвящено очень много работ [16]. Поскольку здесь невозможно дать сколько-нибудь полный обзор данной области исследований, мы рассмотрим только один пример, чтобы получить представление об общей ситуации. Чрезвычайно большие объемы данных, формируемые современными системами спектрозонального зондирования, такими, как система, используемая на борту искусственного спутника Земли ERTS-1 [1]), потребовали проведения интенсивных исследований в области сжатия данных с целью решения проблем, связанных с полосой частот канала передачи, временем обработки данных и емкостью накопителей данных. Поскольку спутниковые данные предназначены для многих экспериментаторов, имеющих различные критерии полезности данных, техника сжатия должна обеспечивать максимальное сохранение передаваемых данных. Другими словами, при сжатии данные не должны разрушаться сверх установленной нормы. Допустимая степень сжатия спектрозональных данных пропорциональна избыточности, т. е. степени корреляции, существующей между элементами данных. Поскольку элементы взаимосвязаны пространственно (в каждом спектрозональном диапазоне существует корреляция яркостей соседних отсчетов изображения) и спектрально (в точке каждого отсчета существует корреляция между результатами четырех спектрозональных измерений), может выполняться пространственное, спектральное или пространственно-спектральное сжатие. Шум вызывает ослабление корре-

[1]) Сканер ИСЗ ERTS способен формировать каждые 20 с один кадр изображения с 3240 × 3240 элементами в каждом из четырех спектрозональных диапазонов. Каждый кадр охватывает приблизительно 185 × 185 км земной поверхности, причем его формирование повторяется с периодом 18 дней.

ляционных связей обоих типов и, следовательно, уменьшает возможности сжатия данных.

С корреляцией тесно связано понятие исходной пригодности данных, пораженных шумом, например, при проведении спектрозональной классификации. В этой процедуре желательно уметь устанавливать класс (например, класс веществ, таких, как кукуруза, вода, песок и т.д.), к которому относится каждый отсчет. Классификация производится путем сравнения вектора измерений данного отсчета со статистиками ряда векторов известных веществ, представляющих все возможные классы; используя один из нескольких возможных методов принятия решения, определяют тот известный вектор, с которым лучше всего согласуется полученный вектор. Точность процедуры классификации наивысшая в том случае, когда известные классы обладают компактными статистиками, а соответствующие им векторы относительно свободно расположены в векторном пространстве. Добавляемый шум вызывает расширение статистик известных классов с соответствующим уменьшением отношения разнос/разброс; кроме того, возникает неопределенность положения вектора неизвестного вещества, подлежащего классификации. Этот



Фиг. 6.1. Влияние случайного шума на точность классификации при спектрозональном зондировании [17]: ——— образцы для обучения; —·— контрольные образцы; — — — пшеница (испытания); ..... соевые бобы (испытания).

эффект исследовали Реди и др. [17]. На фиг. 6.1, заимствованной из работы [17], видно, что с увеличением шума точность классификации падает.

6.1.3. Дополнительные соображения

Фотографические системы и большинство телевизионных камер имеют нелинейную амплитудную характеристику, поэтому действие шума проявляется различно и зависит от яркости участка изображения, на который наложен этот шум [18]. Поскольку эффективная зернистость или шум является функцией размера элемента изображения, проблема зернистости тесно связана с проблемой разрешающей способности [8, 19—24].

В некоторых примерах обработки изображений (такие примеры приведены Оппенгеймом и др. [25] и рассмотрены Шелтоном [26]) могла использоваться шкала квантования, нелинейная по прозрачности, но линейная по плотности; могла также использоваться шкала с каким-то другим распределением уровней. Однако измерительный прибор регистрирует световой поток, который находится в линейном соотношении с прозрачностью фотопленки. Следовательно, именно прозрачность является наиболее важным параметром фотопленки (хотя могут использоваться, конечно, и другие параметры). Таким образом, при рассмотрении шума следует говорить о шуме по прозрачности, а не о шуме по плотности, который имеет значение при изучении визуального эффекта [8].

6.2. Влияние шума на квантование

Следует различать и правильно использовать три разные меры отношения сигнал/шум. Для них мы введем следующие обозначения:

C/σ — отношение полного размаха сигнала (т. е. размаха сигнала от уровня черного до уровня белого) к среднеквадратичному значению шума;

С/Ш — отношение фактического (неполного) размаха сигнала к среднеквадратичному значению шума;

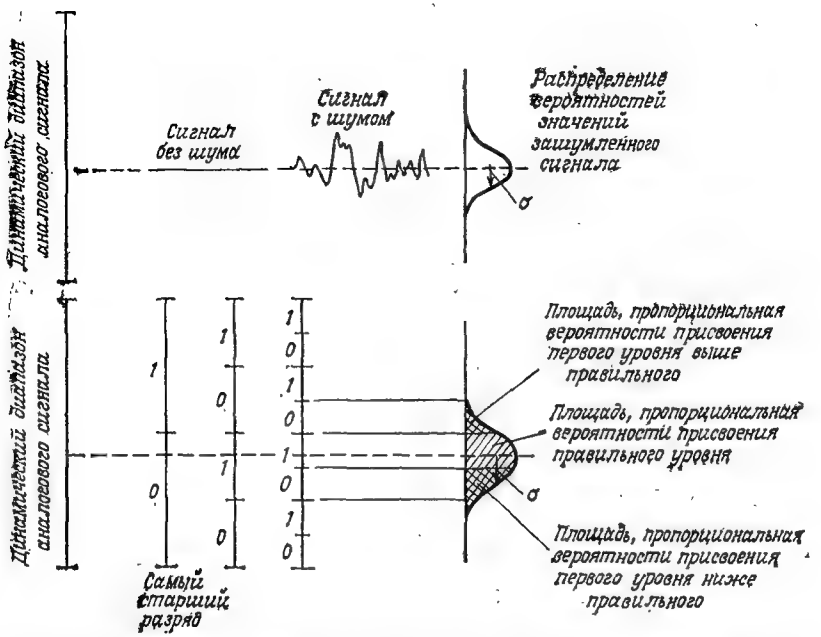
β — отношение шага квантования к среднеквадратичному значению шума.

Здесь «сигнал» — это мера интенсивности света, отсчитываемая от нуля, соответствующего черному, или (в случае использования сканера) прозрачности фотопленки, отсчитываемая от нуля, соответствующего непрозрачности. Таким образом, речь идет об одностороннем сигнале с опорным уровнем черного. В общем случае система должна допускать обработку сигнала

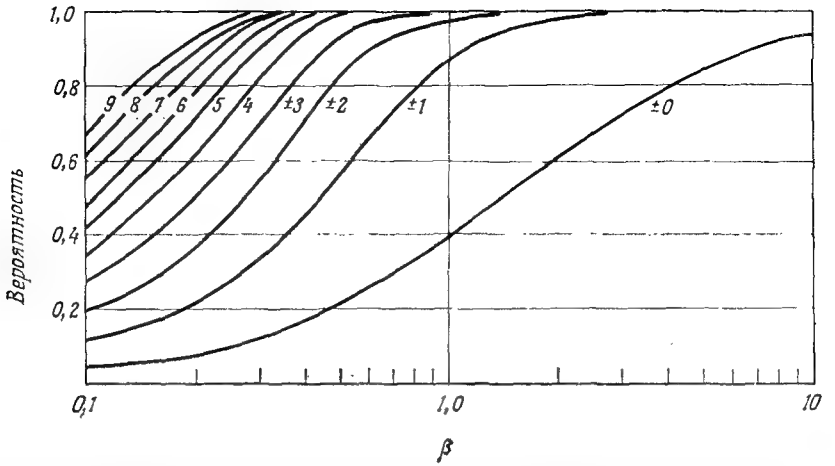
любого уровня в пределах фиксированного и заранее определенного диапазона (например, диапазона 0—1 при измерении прозрачности фотопленки). Поэтому не разрешается произвольно изменять усиление и вводить смещение уровня при использовании сигналов, занимающих лишь часть амплитудного диапазона. Динамический диапазон сигналов, подлежащих квантованию, фиксирован и ограничен.

6.2.1. Квантование сигнала с шумом

Если говорить кратко, назначение квантователя состоит в преобразовании амплитудных изменений сигнала в дискретную совокупность уровней. В системе без шума обозначение любого данного уровня сигнала с помощью соответствующего двоичного числа (ДЧ) не связано с какой-либо неопределенностью. Однако при наличии шума (принимается случайный характер шума) квантуется смесь сигнала с шумом. В этом случае опознавание двоичного числа определяет соответствующий уровень собственно сигнала с некоторой неопределенностью. Выражаясь более конкретно, существует конечная вероятность (определяемая



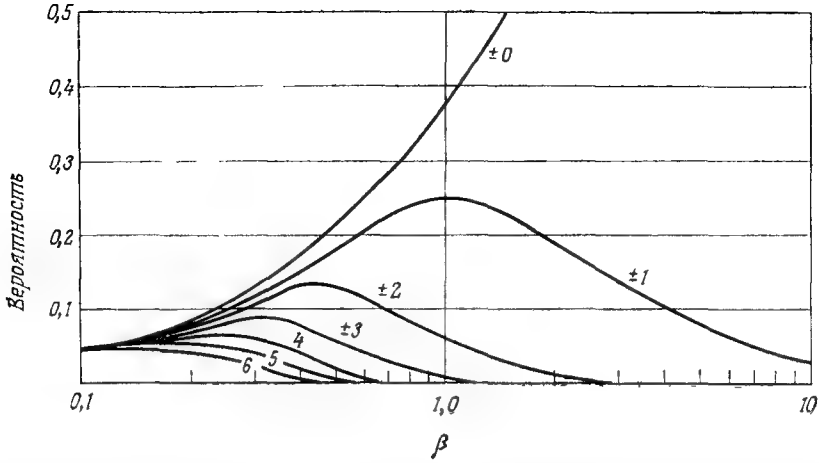
Фиг. 6.2. Поражение шумом «чистого» сигнала приводит к появлению вероятности присвоения неправильного уровня квантования в процессе аналого-цифрового преобразования.



Фиг. 6.3. Дан сигнал, все значения которого равновероятны в пределах диапазона квантования. Дан гауссов шум со среднеквадратичным значением σ. Кривые показывают вероятность правильного присвоения двоичного числа, соответствующего незашумленному сигналу, с ошибкой в пределах цифрового интервала ±0, ±1, ..., ±9 ДЧ (включительно) в зависимости от отношения $\beta =$ (шаг квантования)/σ.

относительными площадями под кривой распределения вероятностей в пределах границ шагов квантования) приписывания неправильного ДЧ данному измеренному уровню. Это иллюстрирует фиг. 6.2. Количественную оценку влияния шума на квантование дал Фридман [27]. Построенная им кривая представлена на фиг. 6.3 как кривая ±0. Эта кривая получена при допущении, что уровень сигнала может принимать с равной вероятностью любое значение в пределах диапазона квантования и что добавляемый шум является гауссовым со среднеквадратичным отклонением σ. Распределение вероятностей смеси сигнала с шумом было получено путем свертки распределения вероятностей сигнала с распределением вероятностей шума. После этого была найдена вероятность правильного квантования как результат интегрирования полученного распределения вероятностей в соответствующих пределах, представляющих границы шага квантования.

В этой связи возникает интересное предположение, касающееся полезности повторного выделения отсчетов. Рассмотрим сначала ситуацию, когда шум отсутствует. В этом случае уровень сигнала имеет строго одинаковые значения в любом из моментов выделения отсчета. Такому сигналу всегда будет приписываться одно и то же ДЧ, что исключает возможность определения точного значения сигнала путем опознавания этого ДЧ.



Фиг. 6.4. Дан сигнал, все значения которого равновероятны в пределах диапазона квантования. Дан гауссов шум со среднеквадратичным значением σ. Кривые показывают вероятность присвоения данного двоичного числа правильному уровню (± 0), уровню выше правильного (+ 1), уровню ниже правильного (—1), ..., (+6), (—6) ДЧ в зависимости от отношения β = (шаг квантования)/σ.

Можно только сказать, что точный уровень сигнала расположен в пределах амплитудного интервала, соответствующего указанному шагу квантования. Однако при наличии шума (даже слабого) появляется конечная вероятность того, что мгновенное значение смеси сигнала с шумом будет несколько, а возможно, и существенно отличаться от мгновенного значения собственно сигнала. Поэтому появляется конечная вероятность того, что новое ДЧ будет отличаться от ДЧ, приписанного «чистому» сигналу. Следовательно, если выделить большое число отсчетов и усреднить полученные ДЧ, то дробная часть усредненного ДЧ укажет на положение истинного уровня сигнала в пределах амплитудного интервала этого ДЧ. Не является ли это указанием на то, что, возможно, лучше пользоваться зашумленным сигналом, чем сигналом без шума?

Фульц [28] сделал еще один шаг вперед. Он расширил рамки анализа, определив вероятность присвоения данного ДЧ уровню, отличающемуся на $\pm L$ ДЧ от правильного уровня. Такая вероятность дается выражением

$$PC_L = (L+1)\operatorname{erf}(L+1)\alpha - L\operatorname{erf} L\alpha - \\ - \{\exp(-L^2\alpha^2) - \exp[-(L+1)^2\alpha^2]\}/\alpha\sqrt{\pi}, \qquad (6.1)$$

где α — отношение шага квантования к $\sigma\sqrt{2}$, а L — уровень квантования, отсчитываемый относительно правильного уровня.

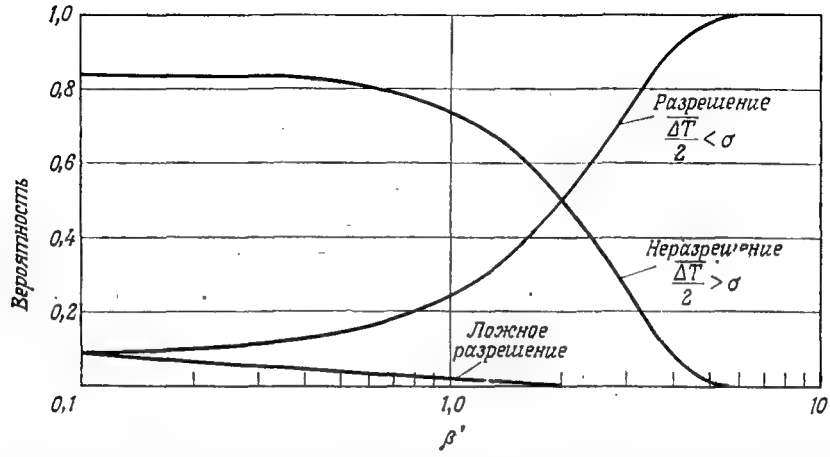
При $L = 0$ вероятность (6.1) становится вероятностью правильного присвоения ДЧ. Кривые вероятности для различных L от 0 до 9, построенные для удобства в функции $\beta = \sqrt{2}\,\alpha = (\text{шаг квантования})/\sigma$, представлены на фиг. 6.3.

Вероятность присвоения конкретному уровню данного ДЧ можно определить, последовательно вычитая из значений кривой для этого уровня на фиг. 6.3 значения кривой, соответствующей ближайшему меньшему уровню. Кривые, полученные в результате вычитаний, показаны на фиг. 6.4.

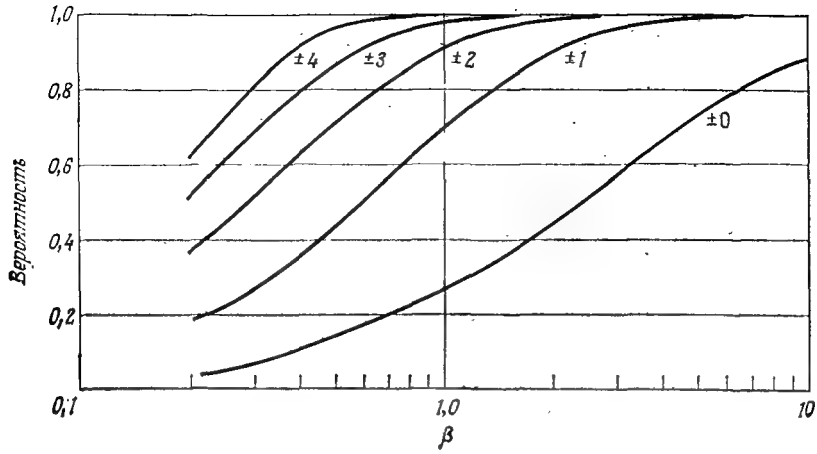
6.2.2. Обнаружимость двух участков различной яркости

Оценим обнаружимость аналитическими средствами двух участков различной яркости или различной прозрачности (для фотопленки). Сначала рассмотрим непрерывный случай без квантования, характеризуемый выполнением аналоговых измерений. Затем рассмотрим случай, когда выделяются и сравниваются квантованные отсчеты.

В случае когда проводятся аналоговые измерения, примем, что среднее различие в прозрачности составляет $\overline{\Delta T}$ и на каждый уровень наложен гауссов шум со среднеквадратичным значением σ. Эквивалентная ситуация возникает при измерении прозрачности двух «однородных», но слегка зашумленных участков. Считывая показания измерителя прозрачности, пытаются



Фиг. 6.5. Даны два участка фотопленки, средние прозрачности которых различаются на $\overline{\Delta T}$. На каждом участке имеется гауссов шум со среднеквадратичным значением σ. Кривые показывают вероятность обнаружения различия между ними (разрешения), неразрешения и ложного разрешения в зависимости от отношения $\beta' = \overline{\Delta T}/\sigma$.



Фиг. 6.6. Даны два сигнала, пораженные гауссовым шумом со среднеквадратичным значением σ. Каждый из них квантуется с одинаковым числом разрядов. Кривые показывают вероятность правильного обнаружения истинной разности двух уровней с ошибкой в пределах цифрового интервала ± 0, ± 1, ..., ..., ± 4 ДЧ (включительно) в зависимости от отношения $\beta =$ (шаг квантования)/σ.

определить, различаются (разрешаются) участки или нет. Можно ввести следующие определения: два участка «разрешаются», если $\overline{\Delta T}/2 > \sigma$, «не разрешаются», если $-\sigma < \overline{\Delta T}/2 < +\sigma$, и «ложно разрешаются» (т. е. опознаются как различные, хотя в действительности различие между ними отсутствует), если $-\infty < \overline{\Delta T}/2 < -\sigma$. Вероятности для всех трех случаев определяются выражениями

$$P \text{ (разрешения)} = [1 + \operatorname{erf}(\alpha'/\sqrt{2} - 1)]/2, \qquad (6.2а)$$

$$P \text{ (неразрешения)} = [\operatorname{erf}(\alpha'/\sqrt{2} + 1) - \operatorname{erf}(\alpha'/\sqrt{2} - 1)]/2, \qquad (6.2б)$$

$$P \text{ (ложного разрешения)} = [1 - \operatorname{erf}(\alpha'/\sqrt{2} + 1)]/2, \qquad (6.2в)$$

где $\alpha' = \overline{\Delta T}/\sqrt{2}\,\sigma$.

Эти вероятности представлены на фиг. 6.5 в виде кривых, построенных в функции $\beta' = \sqrt{2}\,\alpha' = \overline{\Delta T}/\sigma$. Следует отметить, что кривые соответствуют непрерывному случаю и не учитывают эффект квантования.

В эквивалентном случае с квантованием, когда каждый уровень с наложенным шумом квантуется независимо, следует пользоваться кривыми распределения вероятностей на фиг. 6.4. На фиг. 6.6 представлена вероятность правильного обнаружения

разности средних прозрачностей или яркостей двух уровней с ошибкой в пределах цифрового интервала от ± 0 до ± 4 ДЧ.

Подготовленный математический аппарат позволяет приступить к анализу цифровых характеристик системы с шумом.

Анализ проводится на основе 8-разрядной системы, причем используется несколько различных определений понятия «8-разрядная система».

6.2.3. Установление меры качества работы

Следует начать с определения формулировки «квантование с M полезными разрядами». Эта формулировка означает, что в результате действия шума вероятность правильного формирования разряда $M+1$ составляет 0,5. Кривая ± 0 на фиг. 6.3 показывает, что для разряда $M+1$ требуется иметь $\beta = 1{,}4$ (β — отношение шага квантования к среднеквадратичному значению шума); при этом для M-го разряда получаем $\beta = 2{,}8$.

Таким образом, в 8-разрядной системе значение 9-го разряда должно составлять $1{,}4\sigma$ или

$$C/\sigma = 2^{(M+1)} \cdot \beta = 2^9 \cdot 1{,}4 = 717.$$

При таком отношении C/σ вероятность правильного формирования 8-го разряда (для которого $\beta = 2{,}8$) составит $P = 0{,}72$. В целом все результаты измерений не будут отличаться от правильных более чем на ± 2 девятиразрядных ДЧ; с вероятностью $P = 0{,}96$ отличие не превысит ± 1 девятиразрядное ДЧ. Вероятность правильного измерения разности уровней двух участков равна 0,34; $P(\pm 1) = 0{,}8$ и $P(\pm 2) = 0{,}96$.

Если пользоваться тем критерием, что два уровня, яркости которых различаются на 1 ДЧ, должны разрешаться с вероятностью 50%, то потребуется иметь $\beta = 2{,}4$; это соответствует отношению $C/\sigma = 1215$.

Качество работы системы можно также оценивать по уровню шума, вводимого квантователями. Мощность шума, вносимого процессом квантования, определится как [29]

$\sigma_Q^2 =$ (мощность шага квантования)/12 = (шаг квантования)^2/12.

Будем называть систему уравновешенной, если уровень шума квантования равен среднеквадратичной сумме всех шумовых сигналов любого происхождения, генерируемых в системе до квантования. Таким образом, имеем

$$\sigma_{\text{до квантования}} = (\text{шаг квантования})/\sqrt{12}, \qquad (6.3)$$

или

$$\beta = \sqrt{12} = 3{,}56,$$

что дает $C/\sigma = 256 \cdot 3{,}56 = 915$.

В результате рассмотрения всех трех критериев получены требуемые значения β до квантования, равные 2,8, 2,4 и 3,5. Мы будем пользоваться общим эмпирическим правилом, согласно которому на входе квантователя должно обеспечиваться среднеквадратичное значение шума $\sigma \leqslant {}^1/_3 \cdot$ (шаг квантования), что соответствует $\beta = 3$. Это значение достаточно хорошо согласуется со значениями, полученными для всех трех критериев.

6.2.4. Шум системы

Рассмотрим требования к шуму в случае квантования фотографического изображения. Требуемое отношение черно-белого сигнала к среднеквадратичному значению шума (С/σ) можно определить следующим образом:

$$\sigma = {}^1/_3 \cdot (\text{шаг квантования}) = {}^1/_3 \cdot \Delta T \quad (\beta = 3),$$

где $\Delta T = (\text{шаг квантования}) = 1/256$ для 8-разрядной системы, $0 < T < 1$;

$$(\text{размах черно-белого сигнала}) = 1.$$

Плотность и прозрачность на первом уровне после уровня $T=0$ составляют

$$T_1 = \Delta T = 1/256 = 1/N,$$

где N — число уровней квантования,

$$D_1 = \log (1/T_1),$$

где D_1 — плотность фотопленки на первом уровне после $T = 0$. Поэтому

$$\text{С}/\sigma = 1/({}^1/_3 \cdot 1/N) = 3N,$$

С/σ = (черно-белый сигнал)/(среднеквадратичное значение шума).

С/σ в дБ определяется выражением

$$\text{С}/\sigma\,|_{\text{дБ}} = 20 \log (\text{С}/\sigma) = 20 (\log N + \log 3). \qquad (6.4\text{а})$$

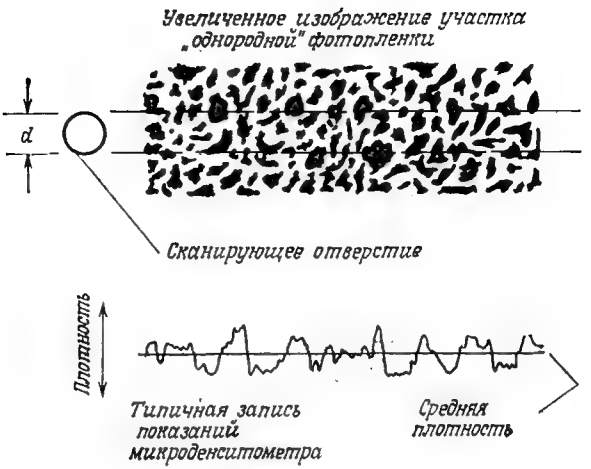
Поскольку $T_1 = 1/N$ и $D_1 = \log (1/T_1) = \log N$, получаем

$$\text{С}/\sigma\,|_{\text{дБ}} = 20 (D_1 + 0{,}477). \qquad (6.4\text{б})$$

В 8-разрядной системе $T_1 = 0{,}039$ и $D_1 = 3{,}4$. Следовательно, С/σ = 77,6 дБ.

6.2.5. Шум зернистости фотопленки

Изображение на обычной галогеносеребряной фотопленке формируется в результате группирования большого числа непрозрачных зерен серебра. Макроскопическим эффектом является образование некоторой средней плотности, определяемой



Фиг. 6.7. Зернистость фотопленки, измеряемая по записи показаний микроденситометра, обычно выражается через среднеквадратичное значение статистической флуктуации плотности σ_D при диаметре сканирующего отверстия 48 мкм и плотности $D \approx 1$.

локальной экспозицией. Однако при последовательном уменьшении диафрагмы с целью достижения все большей и большей разрешающей способности становятся заметными зерна серебра (имеющие вид черных пятен неправильной формы на прозрачном фоне). Это и есть зернистость фотопленки, которая обычно выражается через среднеквадратичное значение статистической флуктуации плотности σ_D, измеренной с помощью отверстия диаметром 48 мкм при плотности $D \approx 1$ (фиг. 6.7).

Как показывает исследование вероятности наложения отдельных серебряных зерен, шум σ_T при произвольной прозрачности T можно определить по известному значению шума σ_{T_1} при прозрачности T_1 по формуле [30]

$$\sigma_T = \sigma_{T_1} \sqrt{\frac{T}{T_1} \frac{(1-T)}{(1-T_1)}}. \tag{6.5}$$

Согласно этой формуле, максимальное значение σ_T достигается при $T = 0{,}5$. Это значение шума представляет наихудший случай. Наибольшее значение шума в наихудшем случае выбирается таким, чтобы обеспечивалось $\beta = 3$ (как и выше). Следовательно, имеем

$$\sigma_{T_{\text{макс}}} = {}^1/_3 \text{(шаг квантования)}.$$

При 8-разрядном квантовании в диапазоне прозрачностей $0 < < T < 1$ получаем

$$\sigma_1 = (1/3 \cdot 256)\sqrt{2T\frac{(1-T)}{(1-1/2)}} = 0{,}0026\sqrt{T(1-T)}.$$

«Стандартные условия», при которых оценивается зернистость, приблизительно соответствуют $T = 0{,}1$ ($D = 1$). Эта замена дает

$$\sigma_{T=0,1} \approx 0{,}00078.$$

Это значение шума σ_T, в наихудшем случае допустимое при $T = = 0{,}1$. Это значение необходимо преобразовать в σ_D (т. е. в σ, измеренное в единицах плотности), поскольку в стандартных технических условиях указывается σ_D. Для плотностей, намного превышающих плотность вуали [30], имеем

$$\sigma_T/T \approx \sigma_D/0{,}4343, \tag{6.6}$$

где σ_D — среднеквадратичное значение зернистости в единицах плотности,

$$\sigma_D = 0{,}4343\sigma_T/T = 0{,}0034.$$

Это наибольшее допустимое значение при $T = 0{,}1$ (измеренное при том же диаметре отверстия, который используется при измерении разрешающей способности).

Было установлено [30], однако, что в практически встречающихся случаях, когда шум имеет равномерный спектр во всем диапазоне полезных пространственных частот, можно записать

$$\sigma_D\sqrt{A} = \text{const} \tag{6.7}$$

(диаметр отверстия много больше 10 диаметров зерна), где A — площадь отверстия. Пересчет на различные диаметры выполняется согласно соотношению

$$\sigma_{D_1}/\sigma_{D_2} = SA_2/SA_1, \tag{6.8}$$

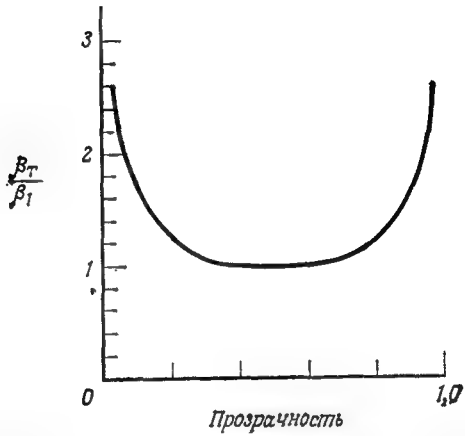
где SA — диаметр сканирующего отверстия в микрометрах. Поэтому

$$\sigma_{D_{48\text{ мкм}}} = \sigma_{D_{SA}}(SA/48) = 0{,}0034 \cdot SA/48 \approx 0{,}00007SA.$$

Таким образом, при данном критерии получаем допустимую зернистость

$$\sigma_D \leqslant 0{,}00007 \cdot SA, \tag{6.9}$$

где σ_D — среднеквадратичное значение зернистости по плотности при $D = 1$ и диаметре отверстия 48 мкм, обеспечивающее $\beta = 3$ в случае 8-разрядного квантования.



Фиг. 6.8. Изменение β, обусловленное зернистостью, в зависимости от прозрачности фотопленки.

Фирма «Кодак» использует значение допустимой зернистости, в 1000 раз превышающее значение зернистости (6.9), получаемое при $SA = 48$ мкм. Так, при использовании типичного сканирующего отверстия диаметром 100 мкм допустимая зернистость $\sigma_D = 0{,}07 \cdot 100 = 7$. Внимательное изучение параметров фотопленок (табл. 6.1) показывает, что зернистость обычной

Таблица 6.1

Типичные значения зернистости некоторых фотопленок фирмы «Кодак» (данные заимствованы из различных проспектов фирмы)

Название	Номер	Зернистость (среднеквадратичное значение $\sigma_D \times 1000$)
Extended red (700)	2457, 2479	32
Extended red (700)	2496	22
Panchromatic high speed	2484	30
Panchromatic	2498	28 (нег.)
		17 (обрат.)
High speed IR (900)	2481, 2424	38
Kodachrome II movie		9
Ektachrome EF	2241	13
Ektachrome MS aerographic	2448	12
Aerocolor neg.	2445	13
Ektachrome IR		17
Aerochrome IR	2443	17

Примечание. Зернистость измеряется при использовании сканирующего отверстия диаметром 48 мкм, объектива с относительным отверстием $F/2{,}0$ и плотности, приблизительно равной 1,0.

пленки намного превышает указанное значение. Следовательно, ограничивающим фактором в системе является качество фотопленки. Поэтому желательно, когда это возможно, производить прямую регистрацию светового изображения с помощью телевизионной камеры (при условии, что камера обладает достаточно высокими параметрами), а не использовать фотопленку в качестве промежуточной среды.

Определим зависимость β от T. Имеем

$$\sigma_T = 1/\beta_T N,$$
$$\sigma_T/\sigma_{T_1} = \sqrt{\frac{T}{T_1}\frac{(1-T)}{(1-T_1)}} = \beta_1/\beta_T. \tag{6.10}$$

Наименьшее значение β (соответствующее наихудшему случаю) получается при $T = 0{,}5$. Обозначая это значение β_1, получаем

$$\beta_T = 0{,}5\beta_1\sqrt{T(1-T)}. \tag{6.11}$$

Соответствующая кривая показана на фиг. 6.8.

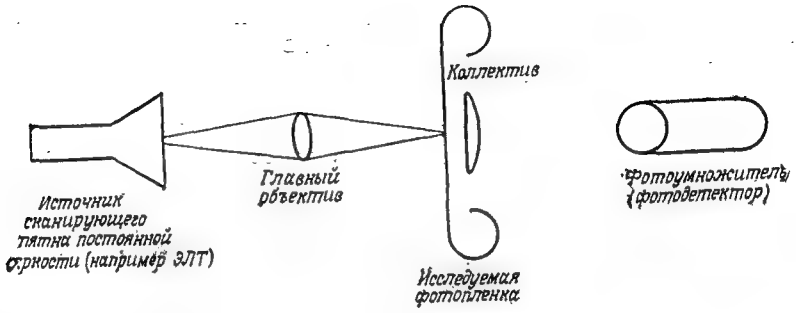
6.3. Полоса частот системы и способы выделения отсчетов

Шум, возникающий в рассматриваемой системе (в том числе шум регистрирующего прибора), можно понизить, если сократить полосу частот системы с неизбежным снижением частоты дискретизации. Максимально допустимую полосу частот системы можно определить либо на основе постоянного отношения С/Ш для всех прозрачностей, либо на основе постоянного отношения C/σ для всех прозрачностей. Выбор между этими двумя возможностями определяется областью применения системы и оценкой, которую исследователь дает различающимся результатам (особенно при малых уровнях освещенности).

Мощности шумов различного происхождения, возникающих в системе, суммируются. Однако на разных участках диапазона яркостей (или прозрачностей) эти мощности имеют различное соотношение. Поэтому указанное суммирование не является систематическим, и в каждом отдельном случае суммирования можно использовать весь допуск по шуму.

6.3.1. Шум фотодетектора

Для определения ширины полосы частот системы в зависимости от шума фотодетектора рассмотрим систему (фиг. 6.9), в которой измеряющий световой поток после ослабления при прохождении сквозь измеряемую фотопленку падает на фотоэлектронный умножитель (с интегральной чувствительностью



Фиг. 6.9. Сканер бегущего луча.

фотокатода $\varepsilon < 1$). Производится подсчет освобожденных электронов, причем результат подсчета обладает статистической флуктуацией. С фотокатодом связан электронный умножитель, вносящий мультипликативный шум. Тепловым шумом нагрузочного резистора, который обычно мал по сравнению с шумами другого происхождения, пренебрегаем.

Можно допустить, что число фотоэлектронов, освобожденных за время отсчета, подчиняется гауссову распределению [31] со среднеквадратичным отклонением σ, равным корню квадратному из среднего значения:

$$\sigma_Q = \sqrt{\bar{Q}}. \tag{6.12}$$

Параметры, соответствующие «полностью открытому» отверстию ($T = 1$), обозначим индексом 0. Поэтому

$$\sigma_{Q_0} = \sqrt{\bar{Q}_0}.$$

Поскольку интегральная чувствительность фотокатода $\varepsilon < 1$, число фотонов превышает число фотоэлектронов, и шум фотонов определяется выражением

$$\sigma_\lambda = \sqrt{\bar{\lambda}} = \sqrt{\bar{Q}/\varepsilon} = \sigma_Q/\sqrt{\varepsilon}, \tag{6.13}$$

где λ — число фотонов, освобожденных за время отсчета.

Шум фотонов складывается с шумом фотоэмиссии; результирующий полный шум, создаваемый фотокатодом, определяется выражением

$$\sigma_e = \sqrt{(\sigma_\lambda \cdot \varepsilon)^2 + \sigma_Q^2} = \sqrt{\left(\sigma_Q \sqrt{\varepsilon}\right)^2 + \sigma_Q^2} = \sigma_Q \sqrt{(1+\varepsilon)}. \tag{6.14}$$

При выводе этого выражения принималось, что весь шум создается флуктуациями тока полезного сигнала. Однако любая паразитная засветка фотодетектора вызывает увеличение флуктуационного шума без соответствующего полезного увеличения

тока сигнала, если измеряемым выходным сигналом является полный ток фотокатода. Такая ситуация возникает, например, в ЭЛТ-сканере (фиг. 6.9). Как отметил Эберхардт [31], в случае использования диссектора в качестве фотодетектора паразитная засветка вокруг полезного элемента изображения подавляется. Более того, если требуется обнаружить очень малые объекты, такие, как изображения звезд, диаметр апертуры следует уменьшать точно до размера изображения; это позволяет уменьшить уровень шума, не связанного с полезным сигналом.

Соответственно наши расчеты основаны на допущении о том, что весь генерируемый шум связан с полезным сигналом и что добавочные шумовые составляющие, связанные с яркостным смещением, отсутствуют. Напрашивается интересное предположение: слабый сигнал (т. е. неглубокая модуляция) должен меньше маскироваться шумом, если он возникает на участке малой яркости; на участке высокой яркости он маскируется сильнее.

Соберем все фотоэлектроны, освобожденные за время отсчета:

$$Q = I\tau/q_e, \tag{6.15}$$

где I — ток фотокатода, τ — время отсчета, q_e — заряд электрона, равный $1{,}6 \cdot 10^{-19}$ Кл, Q — число фотоэлектронов, приходящихся на один отсчет.

Связаный с фотокатодом электронный умножитель имеет коэффициент усиления G; он вносит мультипликативный шум $\sqrt{G(G-1)}$ [31].

Таким образом, получаем

$$I_s = q_e Q G^n/T,$$

$$\sigma_s = (q_e G^n/\tau)\sqrt{QG(1+\varepsilon)/(G-1)},$$

или

$$Q = [G(G+1)/(G-1)]\, I_s^2/\sigma_s^2 =$$
$$= [G(1+\varepsilon)/(G-1)]\,(\text{С/Ш})^2 =$$
$$= K\,(\text{С/Ш})^2,$$

где I_s — ток сигнала на выходе фотоумножителя, G — коэффициент усиления одного каскада, n — число каскадов, σ_s — ток шума на выходе фотоумножителя, $K = (1+\varepsilon)\,G/(G-1)$. При этом получаем

$$I_s = q_e (1+\varepsilon)\, I_s^2 G^n G/\tau\sigma_s^2 (G-1), \tag{6.16а}$$

где I_s/σ_s — отношение С/Ш на выходе, а

$$\tau = q_e (1+\varepsilon)\,(\text{С/Ш})^2\, G^{n+1}/I_s (G-1). \tag{6.16б}$$

Множитель $K=(1+\varepsilon)G/(G-1)$ представляет дополнительный заряд, требуемый для повышения отношения С/Ш на выходе, чтобы в точности скомпенсировать влияние шума, создаваемого фотонами и динодным умножением.

Каждый элемент системы (фотопленка, ЭЛТ, усилитель и т. д.) вызывает уменьшение амплитуды высокочастотных составляющих сигнала, т. е. спад АЧХ на высоких частотах. В общем случае совместный эффект всех высокочастотных спадов в различных узлах системы выражается в ограничении ее полосы частот. Высокочастотная часть результирующей АЧХ имеет форму, напоминающую гауссову. Если принять гауссово ограничение АЧХ, то максимальное сквозное отношение С/Ш в системе достигается при [32]

$$B'\tau' = {}^1/_3, \tag{6.17}$$

где B' — ограниченная полоса частот (в МГц), τ' — длительность импульса (в мкс). Однако после исчезновения входного импульса уровень вершины выходного импульса продолжает возрастать, поэтому ошибки по длительности импульса вызывают ошибки по амплитуде. Компромиссное решение, которое не приводит к существенному снижению отношения С/Ш и тем не менее обеспечивает уменьшение чувствительности к изменению τ', состоит в выборе

$$B'\tau' \approx {}^1/_2.$$

Это соотношение будет использоваться в дальнейшем для оценки требуемой полосы частот, поскольку оно связывает длительность импульса с шириной полосы частот (в Гц):

$$B = I_s(G-1)/2q_e(1+\varepsilon)(\text{С/Ш})^2 G^{n+1}. \tag{6.18}$$

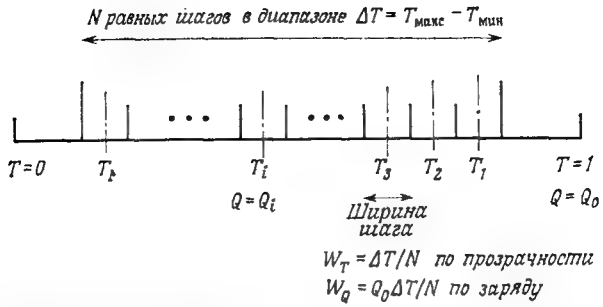
6.3.2. Точность измерений

Используя и развивая метод Шелтона [26], проведем следующие выкладки. Пусть диапазон прозрачностей от $T_{\text{мин}}$ до $T_{\text{макс}}$ содержит N равномерно расположенных уровней (фиг. 6.10). При этом каждому уровню соответствует шаг прозрачности $\Delta T/N$.

Ширина каждого шага, выраженная через Q, составляет

$$W_Q = Q_0 \Delta T/N, \tag{6.19}$$

где $\Delta T = T_{\text{макс}} - T_{\text{мин}}$, Q_0 — значение Q при полностью открытой апертуре. Число собранных электронов должно быть достаточным для обеспечения требуемого $\beta_{\text{макс}}$ при $T_{\text{макс}}$, где $\beta = (\text{шаг заряда})/\sigma_Q = Q_0\Delta T/N\sigma_Q$. На i-м шаге в сторону понижения от $Q_{\text{макс}}$ измеряется число собранных электронов Q_i. Величина Q_0,



Фиг. 6.10. Разделение диапазона прозрачностей от $T_{мин}$ до $T_{макс}$ на N равных шагов [26].

соответствующая открытой апертуре и обеспечивающая получение $\beta_{макс}$ при $T_{макс}$, равна

$$Q_0 = K\beta^2_{макс} N^2 T_{макс}/(\Delta T)^2.$$

Среднее значение T в середине i-го шага в сторону понижения от $T_{макс}$ равно

$$\bar{T}_i = T_{макс} - \Delta T\,(2i-1)/2N.$$

Среднее значение Q на i-м шаге в сторону понижения от $T_{макс}$ равно

$$\bar{Q}_i = Q_0 T_{макс} - W_Q\,(2i-1)/2.$$

На i-м уровне β равно

$$\beta_i = W_Q/\sqrt{\bar{Q}_i K} = \\ = (\Delta T\sqrt{Q_0}/N\sqrt{K})\,[T_{макс} - \Delta T\,(2i-1)/2N]^{-1/2}. \qquad (6.20)$$

Величина β_i в пределах диапазона T, отнесенная к заданной величине $\beta_{макс}$ при $T_{макс}$, равна

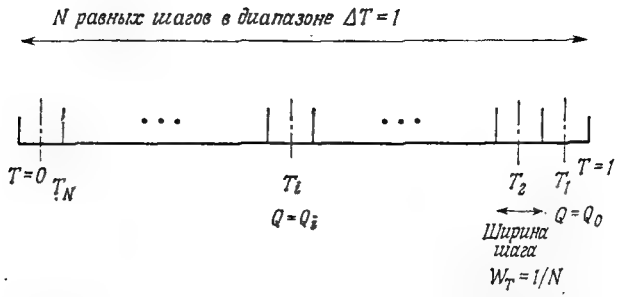
$$\beta_i/\beta_{макс} = \sqrt{T_{макс}}/[T_{макс} - \Delta T\,(2i-1)/2N]^{1/2}. \qquad (6.21)$$

Достижимое значение β_0 при открытой апертуре и при $0 \leqslant T \leqslant 1$ равно

$$\beta_0 = \beta_{макс}\sqrt{T_{макс}}/\Delta T. \qquad (6.22)$$

Отношение С/Ш на любом i-м уровне равно

$$(С/Ш)_i = \sqrt{Q_i}/\sqrt{K} = \\ = [Q_0 T_{макс} - W_Q\,(2i-1)/2]^{1/2}/\sqrt{K} = \\ = \{Q_0\,[T_{макс} - \Delta T\,(2i-1)/2N]\}^{1/2}/\sqrt{K}. \qquad (6.23)$$



Фиг. 6.11. Особый случай $T_{\text{мин}} = 0$ и $T_{\text{макс}} = 1$, встречающийся на практике.

Поэтому отношение $(\text{С/Ш})_i$, отнесенное к отношению $(\text{С/Ш})_{\text{макс}}$ при $T_{\text{макс}}$, равно

$$(\text{С/Ш})_i/(\text{С/Ш})_{\text{макс}} = \{[T_{\text{макс}} - \Delta T\,(2i-1)/2N]/(T_{\text{макс}} - \Delta T/2N)\}^{1/2}. \quad (6.24)$$

Отношение $(\text{С/Ш})_i$, отнесенное к значению отношения С/Ш в наихудшем случае при $i = N$, равно

$$(\text{С/Ш})_i/(\text{С/Ш})_{i=N} = \{[T_{\text{макс}} - \Delta T\,(2i-1)/2N] \times \times [T_{\text{макс}} - \Delta T\,(2N-1)/2N]\}^{1/2}. \quad (6.25)$$

Таким образом, мы видим, что наименьшее значение β — это $\beta_{\text{макс}}$, достигаемое при $T_{\text{макс}}$, а наихудшее отношение С/Ш получается на последнем шаге, который лишь несколько превышает уровень $T = 0$ $(i = N)$.

В особом случае, когда $0 < T < 1$, $\beta_{\text{макс}} = \beta_0$, $T_{\text{макс}} = 1$, $\Delta T = 1$, а отношение $(\text{С/Ш})_{\text{макс}} = \text{С}/\sigma$ (фиг. 6.11).

Случай 1А: увеличение β, начиная с заданного β_0, с уменьшением T:

$$\beta_i/\beta_0\,|_{0<T<1} = [1 - (2i-1)/N]^{-1/2}. \quad (6.26)$$

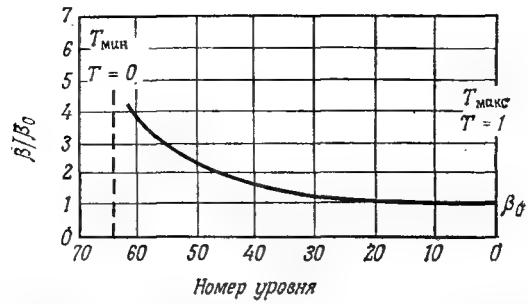
Случай 1Б: уменьшение отношения С/Ш, начиная с отношения С/Ш при $T = 1$, с уменьшением T:

$$(\text{С/Ш})_i/(\text{С}/\sigma)\,|_{0<T<1} = \{[1 - (2i-1)/2N]/(1 - 1/2N)\}^{1/2}. \quad (6.27)$$

Случай 1В: увеличение отношения С/Ш, начиная с заданного отношения С/Ш при $i = N$, с увеличением T:

$$(\text{С/Ш})_i/(\text{С/Ш})_{i=N}\,|_{0<T<1} = \{2N\,[1 - (2i-1)/2N]\}^{1/2}. \quad (6.28)$$

Значение Q, требуемое для обеспечения заданного отношения С/Ш в середине первого шага от нуля, достигается при $i = N$. Это отношение также равно β для этого шага, и в случае



Фиг. 6.12. Увеличение β с понижением прозрачности фотопленки, начиная от β_0 [26].

$0 < T < 1$ получаем для результирующего Q при открытой апертуре выражение

$$Q_0 = 2KN\beta^2, \quad (6.29)$$

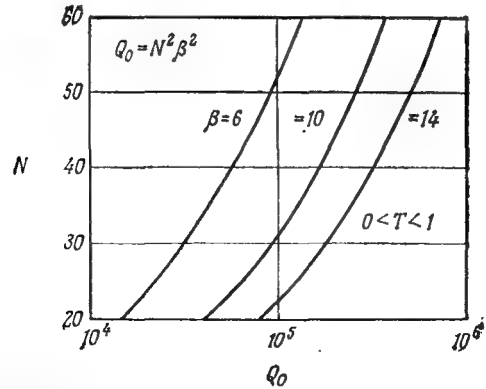
обеспечивающее требуемое β на самом нижнем уровне.

Выражение (6.29) можно сравнить с выражением для Q при открытой апертуре, которое было получено выше:

$$Q_0 = K\beta^2N^2, \quad (6.30)$$

обеспечивающее получение β на самом ярком уровне.

Кривые для β_1/β_0 и N в случае $0 < T < 1$ представлены на фиг. 6.12 и 6.13 соответственно.



Фиг. 6.13. Максимальное число равных шагов прозрачности в зависимости от Q_0 при различных значениях β.

6.4. Практические случаи

Требуемое качество работы системы можно обеспечить в следующих различных режимах:

1. При постоянных (относительно T) параметрах системы, когда
   А) β имеет по крайней мере заданное значение на интервале диапазона T, соответствующем наихудшему случаю, или
   Б) отношение С/Ш имеет по крайней мере заданное значение на интервале диапазона T, соответствующем наихудшему случаю.
2. При переменных (относительно T) параметрах системы, когда
   А) β имеет постоянное значение во всем диапазоне T или
   Б) отношение С/Ш имеет постоянное значение во всем диапазоне T или
   В) используются специальные методы.

Чтобы читатель мог получить представление о реальных цифрах, примем следующие значения параметров:

$$G = 3,\ n = 10,$$
$$I_s = 1 \text{ мА при открытой апертуре,}$$
$$K \approx 2.$$

Это типичные значения параметров фотоэлектронных умножителей, которые практически используются в сканерах с ЭЛТ, имеющей световое пятно малого диаметра.

6.4.1. Случай 1: постоянные параметры системы

Как и прежде, на самом верхнем уровне используем значение $\beta = 3$; диапазон прозрачностей $0 < T < 1$ квантуется с 6 разрядами ($N = 64$).

$Q_0 = 2 \cdot 3^2 \cdot 64^2 = 73\,800$ фотоэлектронов за один отсчет обеспечивают получение заданного β на самом верхнем уровне.

$$\tau = (q_e Q_0 / I_s)\left[G^{n+1}/(G-1)\right] =$$
$$= Q_0 \left(1{,}6 \cdot 10^{-19} \cdot 3^{11}\right)/\left(10^{-3} \cdot 2\right) =$$
$$= 14{,}2 \cdot 10^{-12} Q_0 =$$
$$= 14{,}2 \cdot 10^{-12} \cdot 73{,}8 \cdot 10^3 = 1 \cdot 10^{-6} \text{ с.}$$

$$(\text{С/Ш})_{i=N} = \sqrt{(Q_0/K)\sqrt{(1/2N)}} = 3 \cdot 64/\sqrt{(2 \cdot 64)} = 24\sqrt{2}.$$

Требуемое отношение С/σ системы равно

$$С/\sigma = \beta N = 3 \cdot 64 = 192.$$

Значение С/Ш на самом нижнем уровне меньше значения С/σ на самом верхнем уровне, причем их отношение равно

$$(С/Ш)_{i=N}/(С/\sigma) = (2N-1)^{-1/2} \approx 1/8\sqrt{2} \quad \text{при } 0 < T < 1.$$

Значение β на самом нижнем уровне больше значения β на самом верхнем уровне, причем их отношение равно

$$\beta_{i=N}/\beta_{i=1} = \sqrt{2N} = 8\sqrt{2} \quad \text{при } 0 < T < 1.$$

Если принять, что «мертвое время» между импульсами (за это время происходит перемещение светового пятна, прекращается послесвечение люминофора и производится цифровая регистрация) составляет 20 мкс (практическое значение), то максимальная частота дискретизации (определяемая в основном «мертвым временем») составит приблизительно 50 000 отсчетов в секунду.

Эта частота дискретизации и размеры изображения совместно определяют полное время, необходимое для выделения всех отсчетов изображения. Например, квантование 1024×1024 отсчетов изображения при частоте дискретизации 50 000 отсчетов в секунду занимает приблизительно 20 с.

В аналогичной 8-разрядной системе потребуется увеличить Q в $4^2 = 16$ раз и использовать $\tau = 16$ мкс и $С/\sigma = 768 = 57{,}7$ дБ. И в этом случае приведенные показатели имеют силу для выбранных конкретных параметров.

6.4.2. Случай 2: переменные параметры системы

В некоторых случаях нежелательно или просто невозможно пользоваться постоянными параметрами системы. Например, такая ситуация может возникнуть, когда действующая интенсивность света недостаточна для обеспечения требуемого значения β и поэтому желательно прибегнуть к компенсации путем увеличения времени отсчета (при высокой плотности фотопленки, означающей использование малых шагов прозрачности), или когда желательно иметь как можно больше шагов квантования в пределах всего диапазона прозрачностей, или же когда желательно уменьшить время отсчета до абсолютного минимума. Во всех этих случаях можно использовать факт естественного изменения β или отношения С/Ш, если варьировать некоторые параметры системы в зависимости от яркости или прозрачности, для того чтобы поддерживать постоянным β или отношение С/Ш. Обычно прибегают к варьированию шага квантования (переменный шаг квантования может обеспечить постоянство β или

отношения С/Ш), времени отсчета (например, при сканировании с «постоянной достоверностью» путем интегрирования выходного сигнала для получения постоянного Q в каждой точке отсчета) или яркости сканирующего пятна (при сканировании наиболее плотных участков фотопленки яркость пятна увеличивают).

Применение переменного шага квантования для получения постоянного β

Принимается (как и при выполнении предыдущих вычислений), что фотоэлектроны, испускаемые катодом фотоэлектронного умножителя, имеют распределение Пауссона, которое при больших числах приближается к распределению Гаусса. Среднеквадратичное отклонение σ этого распределения равно корню квадратному из среднего значения. Следовательно, если распределение уровней квантования сделать функцией корня квадратного из T, то на всех уровнях будет обеспечиваться одно и то же значение β. Этот случай рассмотрен Шелтоном [26], который получил выражения для требуемых Q_0 и N. Модифицировав выражения Шелтона с целью учета шума, создаваемого фотонами и динодным умножением, получим выражения

$$\bar{T}_l = (\sqrt{T_{\text{макс}}} - \beta_l/2)^2, \tag{6.31а}$$

$$Q_0 \approx KN^2\beta^2/4\,(\sqrt{T_{\text{макс}}} - \sqrt{T_{\text{мин}}})^2, \tag{6.31б}$$

$$N = 2\sqrt{Q_0}\,(\sqrt{T_{\text{макс}}} - \sqrt{T_{\text{мин}}})/\beta\sqrt{K} \tag{6.31в}$$

при постоянном значении β и переменном шаге квантования. Этот метод позволяет получить больше уровней квантования в заданном диапазоне, чем метод с равными шагами прозрачности или с равными шагами плотности. Недостатком его является введение нелинейности в процедуру цифровых вычислений. Приемлемо это или нет, зависит от требуемой точности вычислений и конкретного содержания изображения. В данном случае невозможно дать общее правило.

Применение переменного шага квантования для получения постоянного отношения С/Ш

Выше был найден способ квантования с постоянным отношением β, заключающийся в изменении шага квантования по определенному закону. Можно попытаться найти способ квантования с постоянным отношением С/Ш. Как и прежде, принимается гауссово распределение фотоэлектронов, среднеквадратичное отклонение σ которого равно корню квадратному из среднего зна-

чения. Таким образом, при любой прозрачности T отношение С/Ш определяется выражением

$$(\text{С/Ш})_T = \sqrt{Q_T}/\sqrt{K} = \text{const}\sqrt{T}. \qquad (6.32)$$

Следует обратить внимание на то, что до введения шагов квантования весь континуум отношений С/Ш полностью определен. Под понятием «квантовать с постоянным отношением С/Ш» можно понимать постепенное изменение шага квантования таким образом, чтобы вероятность правильного квантования (т. е. β) была пропорциональна прозрачности T; при этом отношение $T/\beta = \text{const}$. Поскольку уровень шума увеличивается пропорционально $\sqrt{T}$, шаг квантования должен увеличиваться согласно закону $T^{3/2}$, чтобы отношение β увеличивалось прямо пропорционально T. Такое условие является достаточно произвольным, и поскольку оно не дает видимых преимуществ, мы не будем его больше обсуждать. Обычный метод «постоянной достоверности» обсуждается ниже при рассмотрении случая применения переменных параметров.

Постоянное β при равных шагах квантования

Отношение β можно поддерживать постоянным путем изменения времени отсчета (с соответствующим изменением полосы частот системы) в зависимости от T (или номера уровня квантования i). Начиная с приведенного выше выражения для β_i и используя соотношение $T_i = T_{\text{макс}} - \Delta T(2i-1)/2N$, получаем

$$\beta_i = \frac{\sqrt{Q_0}\,\Delta T}{\sqrt{K}\,N\sqrt{T}}. \qquad (6.33)$$

Как и прежде, полученное выражение относится к системе, в которой измеряемой величиной является Q_i. Можно показать, что β_i постоянно, если отношение Q_0/T является константой, т. е.

$$Q_0 = K\beta^2N^2T/(\Delta T)^2. \qquad (6.34\text{а})$$

В данном случае знак равенства означает «должно быть сделано равным».

При $0 < T < 1$ (6.34а) упрощается и принимает вид

$$Q_0\big|_{0<T<1} = K\beta^2N^2T. \qquad (6.34\text{б})$$

Изменение Q_0 можно обеспечить путем варьирования времени отсчета или яркости сканирующего пятна. Поскольку

$$\tau\, dQ_0/dt,$$

можно варьировать время отсчета τ или скорость накопления заряда dQ_0/dt. В первом случае имеем

$$\tau = T\,(K\beta^2N^2)/(dQ_0/dt) = (\text{масштабный коэффициент}) \cdot T, \quad (6.35)$$

где $K\beta^2N^2/(dQ_0/dt)$ — масштабный коэффициент для всей системы.

Если использовать постоянное τ, то скорость потока фотоэлектронов должна изменяться (путем варьирования яркости сканирующего пятна) согласно соотношению

$$dQ_0/dt = T\,(K\beta^2N^2)/\tau. \quad (6.36)$$

При таком подходе проектирование системы начинается с определения Q_t на уровне квантования с наименьшей прозрачностью T. Используя прежние типичные числовые значения параметров, для $0 < T < 1$ и $N = 64$ получаем

$$Q_0 = 2 \cdot 3^2 \cdot 64^2 \cdot 1/64 = 1152 \text{ фотоэлектрон/отсчет},$$

$$\tau = 14{,}2 \cdot 10^{-12}Q_0 = 16{,}3 \cdot 10^{-9} \text{ с/отсчет}.$$

Время отсчета τ увеличивают прямо пропорционально прозрачности T до значения $\tau_0 = 1$ мкс, соответствующего открытой апертуре ($T = 1$). Благодаря этому используется факт естественного увеличения β при малых T; важность этого фактора возрастает по мере увеличения N. Например, в случае линейного квантования диапазона плотностей 3,0 эквивалентный крайний уровень равен $N = 1024$, так что необходимо иметь $\tau_0 = 256$ мкс для обеспечения $\beta_0 = 3$ при открытой апертуре. Однако при этом получается излишне высокое отношение β на нижних уровнях. Путем преднамеренного снижения величины этого отношения можно сократить время квантования.

Если же получаемое значение τ_0 (например, 1 мкс в приведенном выше примере) достаточно, то можно понижать яркость сканирующего пятна при малых значениях T. В обычной системе при всех условиях устанавливается максимально возможная яркость светового пятна, используемого для сканирования фотопленки. В этом случае возникает ситуация, описанная выше в п. 6.4.1 «постоянные параметры системы». Если сканирующее пятно создает ЭЛТ, то такой режим означает наиболее быстрое старение этой трубки. Поэтому желательно при любой возможности уменьшать яркость сканирующего пятна. Для этого необходимо производить яркостную модуляцию пятна, применив один из возможных способов, подобных описанному ниже.

Реализация этого способа, однако, связана с решением ряда практических проблем. Поскольку τ или яркость пятна должна изменяться пропорционально T, в принципе получается система с положительной обратной связью или система с разомкнутой

петлей и регулированием «вперед». Такая система имеет тенденцию к нестабильности, приводящей к ошибкам измерения. Кроме того, в такой системе требуется, чтобы при $T=0$ величина τ или яркость пятна равнялась нулю. Когда сканирующее пятно попадает на непрозрачную точку или выходит за пределы кадра, яркость падает до нуля и действие системы прекращается. Поэтому требуется принимать специальные меры, чтобы предотвратить это явление. Может оказаться, что желательнее прибегнуть к другим способам использования факта естественного увеличения β для уменьшения времени отсчета.

Постоянное отношение С/Ш *при равных шагах квантования*

В обычных условиях квантование желательно производить таким образом, чтобы β оставалось постоянным: в этом случае обеспечивается постоянная вероятность ошибки, не зависящая от T. Однако при некоторых условиях желательно иметь постоянное отношение С/Ш, не зависящее от T. Такое требование может возникнуть, например, при выполнении нелинейной обработки, какой пользовался Оппенгейм [25], или в случае, когда сканируемое изображение предназначено для наблюдения человеком [26]. Иногда желательно измерить абсолютное значение яркости с заданной точностью. Во всех таких случаях отношение С/Ш (т. е. отношение фактического измеренного размаха сигнала к σ) поддерживается постоянным при изменении T. В результате происходит изменение отношения C/σ в соответствии с изменением T и повышение абсолютной точности по мере приближения T к нулю.

Используя приведенное выше выражение для отношения $(\text{С/Ш})_i$ и соотношение $T_i = T_{\text{макс}} - \Delta T(2i-1)/2N$, получаем

$$(\text{С/Ш})_T = (1/\sqrt{K})\sqrt{\int_0^{\text{порог}} Q_0' T\,dt}\,, \tag{6.37}$$

где Q_0' — скорость потока Q_0.

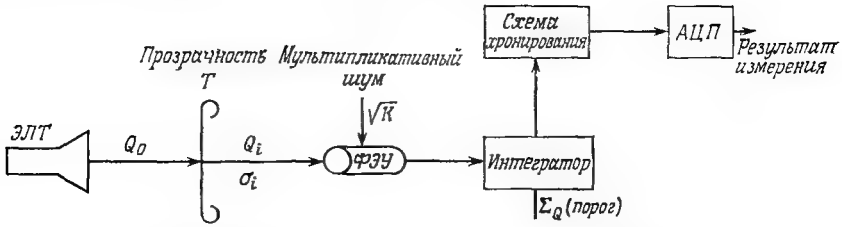
Таким образом, постоянное отношение С/Ш обеспечивается в том случае, когда Q_0' интегрируется до момента получения величины

$$\Sigma_{\text{порог}} = K\,(\text{С/Ш})^2,$$

которая характеризует общее число собранных фотоэлектронов Q_0.

При этом получается переменное β_T, определяемое выражением

$$\beta_T = [(\text{С/Ш})\cdot \Delta T/N]/T. \tag{6.38}$$



Фиг. 6.14. Интегрирование выходного сигнала для получения предопределенного порогового уровня в каждой точке отсчета позволяет использовать время интегрирования в качестве измеряемого параметра.

Обычно Q_0 модулируется путем изменения τ.

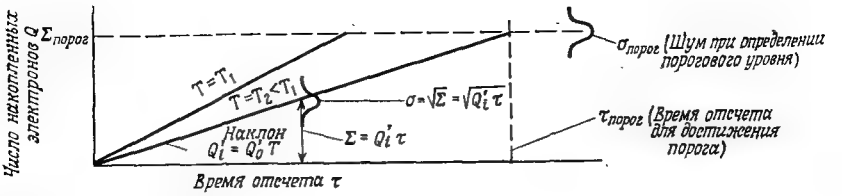
Модуляцию времени отсчета по закону $1/T$ удобно осуществить следующим образом: при передаче любого данного элемента изображения сканирующая апертура задерживается на нем и остается столько времени, сколько требуется для того, чтобы интегрируемый выходной сигнал достиг заранее определенного порогового значения. Соответствующая система сканирования показана на фиг. 6.14.

Интервал времени, требуемый для выполнения такого интегрирования, измеряется. Он изменяется пропорционально обратной величине прозрачности фотопленки или яркости изображения. В такой системе возникает косвенный эффект изменения длительности временного интервала между соседними отсчетами, поэтому при ее проектировании приходится учитывать ряд побочных факторов, таких, как необходимость регистрации несинхронных данных. Рассмотренный подход обычно называют сканированием «с постоянной достоверностью». В данной системе следовало бы варьировать ширину полосы частот в соответствии с модуляцией времени отсчета, чтобы избежать добавления излишнего шума, обусловленного расширением полосы частот сверх требуемой. Однако обычно такая модуляция ширины полосы частот не применяется.

Метод интегрирования с измерением времени интегрирования вносит три новых источника ошибок (т. е. шума).

1. Любая нестабильность процесса определения порогового уровня проявляется как кажущееся колебание уровня выходного сигнала.

2. Искомая прозрачность находится путем вычисления величины, обратной измеренному времени интегрирования. Необходимо принять особые меры для обеспечения достаточно высокой точности измерения временных интервалов, чтобы избежать добавления ошибок усечения и других подобных цифровых погрешностей операции деления.



Фиг. 6.15. К анализу интегрирующего фотодетектора «постоянной достоверности».

3. Точность интегрирования определяется постоянством яркости сканирующего пятна в пределах интервала интегрирования. Поэтому требуется либо предусматривать «абсолютную» стабилизацию яркости пятна, либо использовать дополнительный интегратор для измерения этой яркости с целью последующего нормирования.

Проанализируем процесс основного интегрирования (фиг. 6.15).

Фотоэлектроны Q_i на выходе ФЭУ накапливаются до тех пор, пока их число не достигнет заранее определенной величины $\Sigma_{порог}$.

Скорость накопления равна $Q_0' \cdot T$, где Q' — скорость потока Q.

Полное число накопленных фотоэлектронов за произвольное время τ

$$\Sigma = Q_i'\tau = Q_0'T\tau.$$

Время, требуемое для накопления $\Sigma_{порог}$ фотоэлектронов,

$$\tau_{порог} = \Sigma_{порог}/Q_0'T = (\Sigma_{порог}/Q_0')/T.$$

(Таким образом, время интегрирования обратно пропорционально T.)

Среднеквадратичное отклонение шума в произвольный момент времени τ

$$\sigma_\Sigma = \sqrt{\Sigma} = (KQ_i\tau)^{1/2}.$$

Среднеквадратичное отклонение шума при определении порогового уровня равно $\sigma_{порог}$.

Шум, связанный с обнаружением равенства Σ и $\Sigma_{порог}$ на пороговом уровне, определяется выражением

$$\sigma_\Delta = \sqrt{(\sigma^2_{порог} + \sigma^2_\Sigma)}.$$

Этот шум обусловливает эквивалентную неопределенность измерения τ, зависящую от наклона прямой интегрирования

(т. е. от скорости приближения Σ к $\Sigma_{\text{порог}}$):

$$\sigma_\tau = \sigma_\Delta / Q_i' = \sqrt{(\sigma^2_{\text{порог}} + K\Sigma_{\text{порог}})}\,/Q_i'.$$

Поскольку $Q_i' = Q_0' T$, получаем

$$\sigma_\tau = \sqrt{(\sigma^2_{\text{порог}} + K\Sigma_{\text{порог}})}\,/Q_0' T. \tag{6.39а}$$

При очень малой неопределенности измерения порогового уровня в интеграторе (6.39а) упрощается и принимает вид

$$\sigma_\tau = \sqrt{(K\Sigma_{\text{порог}})}\,/Q_0' T, \tag{6.39б}$$

т. е. обеспечивается постоянное отношение С/Ш:

$$\text{С/Ш} = \tau/\sigma_\tau = \Sigma_{\text{порог}} Q_0' T / Q_0' T \sqrt{(K\Sigma_{\text{порог}})} = \sqrt{(\Sigma_{\text{порог}}/K)}\,. \tag{6.40}$$

Модуляция яркости сканирующего пятна

Модуляцию Q_0 можно также осуществить путем изменения яркости сканирующего пятна. Часто предлагается метод обратной связи, согласно которому яркость пятна увеличивается обратно пропорционально T для обеспечения постоянства выходного светового потока. В этом случае измеряемой переменной величиной является яркость сканирующего пятна. Следует отметить, что этот метод не обеспечивает ни линейного соответствия между шагами Q_0 и T, ни постоянного β, ни постоянного отношения С/Ш. Метод обратной связи характеризуется соотношением

$$Q_i = Q_0 T.$$

При постоянном Q_i (т. е. ограниченном на некотором пороговом уровне) и без учета шумовой составляющей Q_i, внесенной в Q_0 по каналу обратной связи, имеем

$$\left.\begin{array}{c} Q_0 = Q_i/T, \\ Q_0/N < Q_i < Q_0 \\ 1/N < T < 1 \end{array}\right\},$$

$$\beta = (\text{шаг квантования})/\sigma_{Q_0}\sqrt{K} =$$

$$= NQ_0/N\sqrt{K}\,\sqrt{Q_0} = \sqrt{(Q_i/K)}/\sqrt{T}, \tag{6.41а}$$

$$(\text{С/Ш})_{Q_0} = \sqrt{(Q_0/K)} = \sqrt{(Q_i/K\sqrt{T})}\,. \tag{6.41б}$$

Модуляция яркости сканирующего пятна обратно пропорционально T связана со следующей опасностью: при сканировании

наиболее темных участков фотопленки (включая непрозрачные частицы загрязнений) яркость пятна может увеличиться настолько, что в этой точке произойдет прожигание люминофорного экрана ЭЛТ. Для ограничения яркости свечения ЭЛТ до безопасного уровня необходимо вводить специальный ограничитель, а также использовать ту или иную удобную систему сигнализации.

Одна лишь модуляция яркости сканирующего пятна дает мало преимуществ, поскольку время отсчета не варьируется. Если интегратор «постоянной достоверности» охватить обратной связью по свету, то можно обеспечить: линейную зависимость измеряемой переменной τ от T; уменьшение времени отсчета при малых значениях T и, следовательно, среднего времени квантования полного кадра изображения; шумовые характеристики, промежуточные между характеристиками при сканировании с постоянными параметрами и сканировании с постоянным отношением С/Ш.

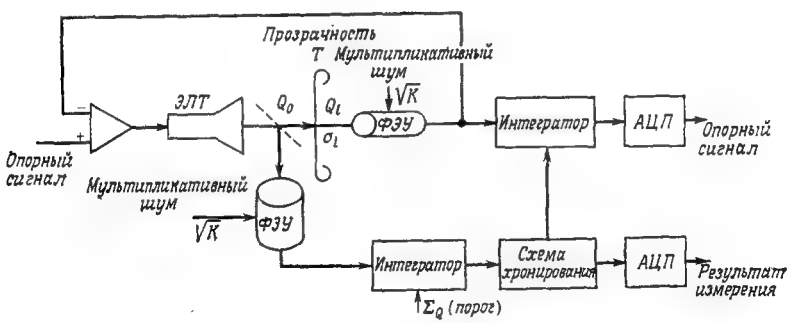
Блок-схема нового сканера показана на фиг. 6.16. Введение обратной связи позволяет поддерживать постоянное значение Q_i. При этом Q_0 увеличивается до значения

$$Q_0 = Q_i/T.$$

Интегратор накапливает Q_0 до момента достижения порогового уровня $\Sigma_{\text{порог}}$:

$$\begin{aligned} \Sigma_Q &= Q_0' \cdot \tau, \\ \tau_{\text{порог}} &= \Sigma_Q/Q_0' = \Sigma_{\text{порог}} T/Q_i'. \end{aligned} \tag{6.42}$$

Таким образом, время выполнения квантования (накопления) линейно зависит от T, т. е. максимально используется запас по качеству при малых T.



Фиг. 6.16. Усовершенствованный сканер с интегрированием и обратной связью.

Обратная связь обеспечивает постоянство Q_i; поэтому

$$Q_0 \approx (R - \sigma_i)/T,$$
$$\sigma_{Q_0} = \sqrt{Q_0} = (\text{масштабный коэффициент}) \cdot 1/\sqrt{T} =$$
$$= f(R - \sigma_i) = \sqrt{(M/T)}, \tag{6.43}$$

где M — модуляционный масштабный коэффициент.

Можно определить полный уровень шума на выходе интегратора (расширяя рамки проведенного анализа):

$$\sigma_{\text{полн}} = T\left(\sigma^2_{\text{порог}} + \sigma_{\text{порог}} Q_i/T + M/T\right)^{1/2}/Q_i =$$
$$= T\left(\sigma^2_{\text{порог}} + \Sigma_{\text{порог}} + M/T\right)^{1/2}/Q_i, \tag{6.44}$$

где $\sigma_{\text{порог}}$ — шум при определении порогового уровня; $\Sigma_{\text{порог}}$ — шум, связанный с накопленными фотоэлектронами $Q_{\text{порог}}$; M/T — модуляционный шум в Q_0, обусловленный действием обратной связи по Q_i.

Таким образом, мы видим, что в зависимости от соотношения между тремя составляющими шума $\sigma_{\text{полн}}$ может быть пропорционально T или $\sqrt{T}$ или может быть некоторой промежуточной функцией T. Характер зависимости определяется конкретными параметрами сканера.

Поскольку

$$\left.\begin{aligned} \tau &= T\Sigma_{\text{порог}}/Q_i', \\ (\text{С/Ш}) &= \tau/\sigma_\tau = (\text{масштабный коэффициент}) \end{aligned}\right\} \text{ при } \sigma_\tau \approx (\text{масштабный коэффициент}) \cdot T, \tag{6.45а}$$

или

$$\left.\begin{aligned} (\text{С/Ш}) &= (\text{масштабный коэффициент}) \cdot \sqrt{T} \\ \beta &= (\text{шаг квантования})/\sigma_\tau = (\text{масштабный коэффициент})/\sqrt{T} \end{aligned}\right\} \text{ при } \sigma_\tau \approx (\text{масштабный коэффициент}) \times \sqrt{T}. \tag{6.45б}$$

Поскольку действие обратной связи может привести к недопустимому увеличению яркости, требуется использовать специальный ограничитель. Вырабатываемый опорный сигнал используется для измерения фактического значения полного Q_i, которое служит для нормирования результата измерения с целью компенсации эффекта ограничения яркости и колебаний Q_i в пределах интервала интегрирования.

Расчет системы следует начинать с наименьших значений β (соответствующих максимальным значениям T). Для этого участка диапазона вычисляют σ и Q_0, пользуясь прежней методикой; при этом в АЦП интегратора следует предусмотреть доста-

точный «запас битов», чтобы можно было производить измерения на краю диапазона с малыми значениями T. Опорный сигнал прямо пропорционален Q_i; его следует масштабировать таким образом, чтобы он содержал составляющую, пропорциональную максимальному значению Q_i, которое достигается при наименьших значениях T.

Данный подход не обеспечивает постоянного значения β, но он обладает рядом преимуществ, перечисляемых ниже.

Преимущества по сравнению с обычными методами сканирования «с постоянной достоверностью»:

1) измеряемая переменная τ линейно связана с T, что позволяет устранить операцию деления;
2) повышенное быстродействие при $T < 0{,}5$ $(D > 0{,}3)$, что соответствует типичному случаю;
3) возможность обеспечения большего постоянства β.

Преимущества по сравнению с обычным методом сканирования с обратной связью:

1) используется возможность увеличения яркости сканирующего пятна с целью уменьшения τ;
2) измеряемая переменная линейно связана с T, что позволяет устранить операцию деления.

Преимущества по сравнению с методом сканирования с постоянными параметрами:

1) гораздо более высокое быстродействие, поскольку в сравниваемой системе постоянный параметр τ рассчитывается для наихудшего случая;
2) повышенный срок службы ЭЛТ, поскольку яркость сканирующего пятна повышается только при передаче участков с малыми значениями T.

6.5. Выводы

Рассмотрение влияния шума на точность квантования (в цифровой системе) позволяет заключить, что удобным параметром для оценки системы является отношение шага квантования к среднеквадратичному значению шума β. Показано далее, что приемлемое значение этого параметра равно трем. Проведен анализ шумов различного происхождения (шум зернистости фотопленки, шум электронов, дробовой шум) с целью определения взаимосвязи между их параметрами и яркостью квантуемого изображения. На основе этого анализа исследованы характеристики ряда стандартных сканирующих систем бегущего луча. Наконец, предложена новая блок-схема усовершенствованного

сканера, включающая многие узлы стандартных систем. Предлагаемая система обладает улучшенными шумовыми характеристиками и уменьшенным временем квантования. Дальнейший анализ должен показать, останется ли оптимальной концепция нового сканера в том случае, если фотоэлектронные умножители заменить на другие фотодетекторы, такие, как кремниевые фотодиоды, а вместо электроннолучевой трубки использовать лазерный источник света.

Литература

1. *Schade O. H., Sr.*, J. Soc. Mot. Pict. Telev. Engrs., **56**, 137 (1951); **58**, 181 (1952); **61**, 97 (1953); **64**, 593 (1955).
2. *Schade O. H., Sr.*, J. Opt. Soc. Am., **46**, 721 (1956).
3. *Stultz K. F., Zweig H. J.*, J. Opt. Soc. Am., **49**, 693 (1959).
4. *Stultz K. F., Zweig H. J.*, J. Opt. Soc. Am., **52**, 45 (1962).
5. *Hacking K.*, J. Brit. IRE, **23**, 307 (1962).
6. *Huang T. S.*, IEEE Trans. Info. Theory, **IT-11**, 43 (1965).
7. *Heyning J. M.*, The Human Observer, Proc. Seminar on Human in the Photo-Opt. Syst., SPIE, New York City, 1966.
8. *Altman J. H.*, J. Soc. Mot. Pict. Telev. Engrs., **76**, 629 (1967).
9. *Roberts L. E.*, IRE Trans. Info. Theory, **IT-8**, 145 (1962).
10. *Higgins G. C.*, Information Capacity of Photographic Materials в книге Photographic Systems for Enginers, Soc. Photo. Sci. Engrg., Washington, D. C., 1966, pp. 167—207.
11. *Doerner E. C.*, J. Opt. Soc. Am., **52**, 669 (1962).
12. *Fry G. A.*, J. Opt. Soc. Am., **53**, 361 (1963).
13. *Schade O. H., Sr.*, J. Soc. Mot. Pict. Telev. Engrs., **73**, 81 (1964).
14. *Blackman E. S.*, Phot. Sci. Eng., **12**, 244 (1968).
15 *Brault W.*, Astron. Astrophys., **13**, 169 (1971).
16. *Wilkins L. C., Wintz P. A.*, Bibliography on Data Compression, Picture Properties and Picture Coding, Purdue University Technical Report TR-EE69-10.
17. *Ready P. J., Wintz P. A., Whitsitt S. J., Landgrebe D. A.*, Effects of Compression and Random Noise on Multispectral Data, Proc. 7th Symposium on Remote Sensing of the Environment, Univ. Michigan, 1971, pp. 1321—1343.
18. *Rindfleisch T. C., Dunne J. A., Frieden H. J., Stromberg W. D., Ruiz R. M.*, J. Geophys., **76**, 394 (1971).
19. *Levi L.*, J. Opt. Soc. Am., **48**, 9 (1958).
20. *Zweig H. J., Higgins G. C., MacAdam D. L.*, J. Opt. Soc. Am., **48**, 926 (1958).
21. *Jones R. C.*, J. Opt. Soc. Am., **51**, 1159 (1961).
22. *Eyer J. A.*, Phot. Sci. Eng., **6**, 71 (1962).
23. *Levi L.*, Phot. Sci. Eng., **7**, 26 (1963).
24. *Hacking K.*, J. Mot. Pict. Telev. Engrs., **73**, 1015 (1964).
25 *Oppenheim A. V., Schafer R. W., Stockham T. G.*, Proc. IEEE, **56**, 1264 (1968). [Имеется перевод: ТИИЭР, **56**, № 8, 5 (1968).]
26. *Shelton C. F., Herd H. H., Leybourne J. J.*, Grey-Level Resolution of Flying Spot Scanner Systems, SPIE Photo Optical Systems Seminar, Rochester, N. Y., 1967.
27. *Friedman H. D.*, Proc. IEEE, **53**, 658 (1965). [Имеется перевод: ТИИЭР, **53**, № 6, 760 (1965).]
28 *Fultz G. L.*, The Effect of Source Noise on Quantization Accuracy and on PE Statistics, JPL Tech. Memo 3341-65-5, 1965.

29. *Walli C. R.*, Quantizing and Sampling Errors in Hybrid Computation, Proc. Fall Joint Computer Conference, 1964, pp. 545—558.
30. *O'Neill E. L.*, Introduction to Statistical Optics, Addison-Wesley, Reading, Mass. Palo Alto, London, 1963, pp. 105—121.
31. *Eberhardt E. H.*, Signal-to-Noise Ratio of Image Dissector, Technical Note No. 101, ITT Industrial Laboratories, Fort Wayne, Indiana, 1966.
32. *Schwartz M.*, Information Transmission, Modulation, and Noise, McGraw-Hill, New York, London, 1959.
33. *Stockham T. G.*, Proc. IEEE, **60**, 828 (1972). [Имеется перевод: ТИИЭР, **60**, № 7, 93 (1972).]

Библиография [1])

Глава 1

Huang T. S., Image Processing, North-Holland, Amsterdam, 1976.

Huang T. S., Barker D., Berger S., Iterative image restoration, Appl. Opt., **14**, 1165 (1975).

Huang T. S., Burnett J., Deczky A., The importance of phase in image processing filters, IEEE Trans. on Acoustics, Speech, and Signal Processing, **ASSP-23**, 529 (1975).

Huang T. S., Narendra P., Image restoration by singular value decomposition, Appl. Opt., **14**, 2213 (1975).

Proc. IEEE, Special issue on Digital Signal Processing, **63**, No. 4 (1975). [Имеется перевод: ТИИЭР, **63**, № 4 (1975).]

IEEE Trans. on Circuits and Systems, Special issue on Digital Filtering and Image Processing, **CAS-22**, No. 3 (1975).

Julien G. A., Sid-Ahmed, Stability constraints used in computeraided design of recursive digital filters, IEEE Trans. Acoustics, Speech and Signal Processing, **ASSP-22**, 153 (1974).

Oppenheim A. V., Schafer R., Digital Signal Processing, Prentice-Hall, New York, 1975.

Rabiner L., Gold B., Theory and Applications of Digital Signal Processing, Prentice-Hall, New York, 1975. (Имеется перевод: Рабинер Л., Голд Б., Теория и применение цифровой обработки сигналов, Мир, М., 1975.)

Rosenfeld A., Kak A. C., Digital Picture Processing, Academic Press, New York, 1976.

*Computer, Special issue on Digital Picture Processing, **7**, No. 5 (1975).

Green W. B.*, Computer Image Processing — the Viking Experience, IEEE Trans. on Consumer Electronics, **CE-23, 281 (1977).

*Proc. IEEE, Special issue on Multidimentional Systems, **65**, No. 6 (1977). [Имеется перевод: ТИИЭР, **65**, № 6 (1977).]

Глава 2

Andrews H. C., Patterson C. L., Outer products and their uses in digital image processing, Am. Math. Monthly, **82**, 1 (1975).

Andrews H. C.*, Monochrome digital image enhancement, Appl. Opt., **15, 495 (1976).

Andrews H. C.*, Singular Value Decompositions and Digital Image Processing, IEEE Trans. on Acoustics, Speech, and Signal Processing, **ASSP-24, 26 (1976).

Reed I. S., Truong T. K., Kwoh Y. S., Hall E. L.*, Image Processing by Transforms Over a Finite Field, IEEE Trans. on Computers, **C-26, 874 (1977).

Hou H. S., Andrews H. C.*, Least Squares Image Restoration Using Spiine Basis Functions, IEEE Trans. on Computers, **C-26, 856 (1977).

---

[1]) Звездочкой отмечена литература, добавленная при переводе.

**Kerke H. B., Solanki J. K.*, Modified Slant Haar Transforms for Image Data Compression, Comput. and Elect. Engng., 4, 199 (1977).
Woods J. W., Radewan C. H.*, Kalman Filtering in Two Dimensions, IEEE Trans. on Information Theory, **IT-23, 473 (1977).
Twogood R. E., Ekstrom M. P.*, An Extention of Eklundh's Matrix Transposition Algorithm and Its Application in Digital Image Processing, IEEE Trans. on Computers, **C-25, 950 (1976).
Pendergrass N. A., Mitra S. K.*, Spectral Transformations for Two-Dimensional Digital Filters, IEEE Trans. on Circuits and Systems, **CAS-23, 26 (1976).
Jain A. K.*, An Operator Factorization Method for Restoration of Blurred Images, IEEE Trans. on Computers, **C-26, 1061 (1977).
Jain A. K.*, Image Coding Via a Nearest Neighbours Image Model, IEEE Trans. Commun., **COM-23, 318 (1975).
**Dail J. E.*, Polynomial Transform Coding of Image Signals, Int. conf. on communications, v. 1 (Conf. Rec.), New York, 1976.

Глава 3

Kamp Y., Thiran J. P., Maximally flat nonrecursive two-dimentional digital filters, IEEE Trans. Circuits and Systems, **CAS-21**, 437 (1974).
Kamp Y., Thiran J. P., Chebyshev approximation for two-dimentional nonrecursive digital filters, IEEE Trans. Circuits and Systems, **CAS-22**, 208 (1975).
Rabiner L. R., McClellan J. H., Parks T. W.*, FIR Digital Filter Design Techniques Using Weighted Chebyshev Approximation, Proc. IEEE, **63, 595 (1975). [Имеется перевод: ТИИЭР, **63**, № 4, 61 (1975).]
Prasad K. P., Reddy P. S.*, Frequency Response of 2-Dimentional Digital Filters, Electronics Letters, **14, 43 (1978).
Mersereau K. P., Mecklenbräuker W. F. G., Quatieri T. F., Jr.*, McClellan Transformations for Two-Dimentional Digital Filtering: I — Design, IEEE Trans. on Circuits and Systems, **CAS-23, 405 (1976).
Mecklenbräuker W. F. G., Mersereau R. M.*, McClellan Transformations for Two-Dimensional Digital Filtering: II — Implementation, IEEE Trans. on Circuits and Systems, **CAS-23, 414 (1976).
Mersereau R. M., Dudgeon D. E.*, Two-Dimentional Digital Filtering, Proc. IEEE, **63, 610 (1975). [Имеется перевод: ТИИЭР, **63**, № 4, 79 (1975).]

Глава 4

Bose N. K., Kamat P. S., Algorithm for stability test of multi-dimensional filters, IEEE Trans. Acoustics, Speech, Signal Proc., **ASSP-22**, No. 5 (1974).
Bose N. K., Jury E. I., Positivity and stability test for multi-dimentional filters (discrete — continuous), IEEE Trans. Acoustics, Speech, Signal Proc., **ASSP-22**, No. 3 (1974).
Costa J. M., Venetsanopoulos A. N., Design of circularly symmetric two-dimentional recursive filters, IEEE Trans. Acoustics, Speech, Signal Proc., **ASSP-22**, No. 6 (1974).
Jury E. I., The theory and applications of the inners, Proc. IEEE, **63**, 1044 (1975). [Имеется перевод: ТИИЭР, **63**, № 7, 60 (1975).]
Maria G. A., Fahmy M. M., On the stability of the two-dimensional digital filters, IEEE Trans. Audio Electroacoustics, **AU-21**, 470 (1973).
Maria G. A., Fahmy M. M., An L_p design technique for two-dimentional digital recursive filters, IEEE Trans. Acoustics, Speech, Signal Proc., **ASSP-22**, No. 1 (1974).
Mersereau R. M., Dudgeon D. E., The representation of two-dimensional sequences as one-dimensional sequences, IEEE Trans. Acoustics, Speech, Signal Proc., **ASSP-22**, No 5 (1974).

Mitra S. K., Sagar A. D., Pendergrass N. A., Realizations of two-dimensional recursive digital filters, IEEE Trans. Circuits and Systems, **CAS-22**, No. 3 (1975).
Ni M.-D., Aggarwal J. K., Two-dimensional digital filtering and its error analysis, IEEE Trans. Computers, **C-23**, No. 9 (1974).
Siljak D. D., Stability for two-variable polynomials, IEEE Trans. Circuits and Systems, **CAS-22**, No. 3 (1975).
Dudgeon D. E.*, Two-Dimensional Recursive Filter Design Using Differentional Correction, IEEE Trans. Acoustics, Speech, Signal Proc., **ASSP-23, 264 (1975).

Глава 5

Aboutalib A. O., Silverman L. M., Restoration of images degraded by curvi-linear motion, in Second International Joint Conference on Pattern Recognition, Copenhagen, 1974, IEEE Cat. No. 74CHO885-4C.
Burke J. J., Estimating objects from their blurred and grainy images, Proc. of the Technical Program Electro-Optical Systems Design Conference (West International Laser Exposition, San Fransisco, 1974).
Frieden B. R., Restoration of Pictures by Monte Carlo allocation of pseudo-grains, in Second International Joint Conference on Pattern Recognition (Copenhagen, 1974, IEEE Cat. No. 74CHO885-4C).
Sabri M. S., Steenaart W.*, An Approach to Band-Limited Signal Extrapolation: The Extrapolation Matrix, IEEE Trans. on Circuits and Systems, **CAS-25, 74 (1978).
Keshavan H. R., Srinath M. D.*, Sequential Estimation Technique for Enhancement of Noisy Images, IEEE Trans. on Computers, **C-26, 971 (1977).
Cannon M.*, Blind Deconvolution of Spatially Invariant Image Blurs with Phase, IEEE Trans. on Acoustics, Speech, Signal Proc., **ASSP-24, 58 (1976).
Honda T., Kumagaya K., Tsujiuchi J.*, Restoration of Defocused Pictures by Image-Scanning method, Optica Acta, **24, 23 (1977).

Глава 6

Chien R. T., Snyder W. E., Hardware for visual image processing, IEEE Trans. Circuits and Systems, **CAS-22**, 541 (1975).
Mannos J. L., Sakrison D. J., The effect of visual criterion on encoding of images, IEEE Trans. Information Theory, **IT-20**, 525 (1974).
Schreiber W. F., The effect of scanning speed on the S/N ratio of camera tubes, Proc. IEEE, **52**, 217 (1964). [Имеется перевод: ТИИЭР, **52**, № 2, 230 (1964).]
Hunt B. R., Breedlove J. R.*, Scan and Display Considerations in Processing Images by Digital Computer, IEEE Trans. on Computers, **C-24, 848 (1975).
Sawchuk A. A.*, Real-Time Correction of Intensity Nonlinearities in Imaging Systems, IEEE Trans. on Computers, **C-26, 34 (1977).
Netravali A. N., Prasada B.*, Adaptive Quantization of Picture Signals Using Spatial Masking, Proc. IEEE, **65, 536 (1977). [Имеется перевод: ТИИЭР, **65**, № 4, 47 (1977).]
Netravali A. N., Prasada B.*, Quantization of Color Signals, Proc. IEEE, **65, 1177 (1977). [Имеется перевод: ТИИЭР, **65**, № 8, 85 (1977).]

Предметный указатель

Алгебраические функции 152
Аподизаторы 214
Аппроксимация рядом разделимых функций 77
Априорная информация 24, 27, 32, 194, 202, 233—240

Билинейное преобразование 157
Блочная тёплицева матрица 74
— фурье-матрица 74
Большие изображения 209
Быстрое преобразование Фурье (БПФ) 85, 119, 120, 131, 206, 209

Векторные внешние произведения 37, 38
Весовая функция 111
Возникновение ложных частот 102
Восстановление по проекциям 196
Второй алгоритм Ремеза 81
Выделение посредством дельта-функции 199
— признаков изображения 33
Выпуклая оболочка 91

Гибридное кодирование 61

Двоичное число (ДЧ) 271
Двойная звезда альфа Лебедя 246
Двумерная инверсия по критерию наименьших квадратов (ДИНК) 174
— передаточная функция 145
Дельта-функция Дирака 198
Дискретная противосвертка 221—224
Дискретное преобразование Фурье (ДПФ) 85, 105, 179
Дискретные системы (двумерные) 82—86
Диссектор 289
Дистанционное исследование (зондирование) 195
Дифференциальная ИКМ (ДИКМ) 53—55, 62
Дифференциаторы (двумерные) 89, 134—137

Единичный импульс 82

Зернистость 276, 284, 285

Идеальная интерполяция 197
— частотно-ограниченная оценка 198
Избыточность 274
Изображение 34, 193, 273
Изопланатизм 207
Импульсная характеристика 34, 83, 146, 169
Инверсия уравнений Фредгольма 196
Инверсная свертка 196
— фильтрация 208—214
Интервал Найквиста 198
Интерполяционная формула Лагранжа 115, 140
Информация по Шеннону 227
ИСЗ ERTS-1 274
Искажение изображения 21, 27
— — аппроксимация рядом разделимых функций 77
— — выражаемое столбцом 36
— — — строкой 36
— — разделимое 36

Квадратичное программирование 26, 27
Квантование 271, 276—287
Кодирование изображений 9, 28, 31, 40, 274
— — гибридное 61
— — дифференциальная ИКМ (ДИКМ) 53—55, 62
— — пространственное 53—56
— посредством преобразований 40—52, 56—63

Корневые диаграммы 149
Косинусное преобразование 60
Кремниевые фотодиоды 306
Кронекерово (прямое) произведение матриц 53, 65

Лагранжиан 226
Лазерный источник света 306
Линейная система 82
— — инвариантная по отношению к сдвигу 82
Линейное программирование 17, 81, 95—100, 105, 112
— — базисные переменные 97, 122
— — вектор ограничений 96
— — — программы 96
— — — стоимости 96
— — двойственная задача 96, 98, 116, 126, 127
— — комплекс программ 100, 110, 114, 117, 127
— — матрица коэффициентов 96
— — нормальные случайные переменные 95
— — ограничения в виде неравенства 95
— — — — — равенства 95
— — основная задача 96
— — переработанный симплексный алгоритм 100
— — произвольные переменные 97
— — свободные переменные 95
— — симплексный алгоритм 96, 138
— — целевая функция 96

Максимумы и минимумы (локальные) двумерных функций 121, 123, 141
Марковский процесс (двумерный) 27, 56, 59
Матрица Шура — Кона 155
Матрица-циркулянт 60
Матричные квадратичные формы 31
Машинная графика 10, 28
Метод Винера — Хопфа 15
— Монте-Карло 261—264
— проекций 21—26
Методы минимизации без ограничений 16
— оптимизации 16, 26
Микроденситометр 273
Минимальная фаза 174
Многомерные сигналы 9
Модели изображения 33—36
Модуляционная передаточная функция (МПФ) 212, 273

Наилучшая среднеквадратичная оценка 206
Наклонное преобразование 61
Нерекурсивные цифровые фильтры (двумерные) 15, 83

Обобщенный полином 93, 115
Обострение характеристик антенн 196
Обработка изображений 9
Ограничения по симметрии импульсной характеристики (двумерной) 87, 88, 111, 119, 130, 132, 134
Оптическая передаточная функция (ОПФ) 273
Оригинал (объект) 34, 194, 273
Ортогональные базисные функции 39
Отношение сигнал/шум (С/Ш) 272, 276, 289
Оценка главного значения 198
— плотности вероятности по моментам 196, 249
— по максимуму правдоподобия 257
— — — энтропии 248—261
— энергетического спектра 196
Ошибка отбора 97, 128—131

Пакетные операции 53, 65, 74
Передаточная функция 84
Переработанный симплексный алгоритм 100
Переходная полоса 108
Повторное выделение отсчетов 278
Подавление постоянной составляющей 215, 216
Полоса задерживания 108
— пропускания 108
— частот 287, 290, 297, 300
Полосовые фильтры (двумерные) 88
Преобразование Адамара 41—46, 61
— в дискретном линейном базисе 61
— Ганкеля 101
— Гильберта (двумерное дискретное) 178, 179
— Карунена — Лоэва 40, 58, 59
— Лапласа (двумерное) 156
— тождественности 42
— Фурье (ПФ) 28, 46, 60, 196, 273
— — интерферограмм 195
— Хаара 42—47, 61
— Хотеллинга 58
Приборы на поверхностных акустических волнах 19
— с зарядовой связью (ПЗС) 28
Проверка Гурвица 154

Проектирование нерекурсивных цифровых фильтров (двумерных), вычислительная сложность 126, 138
— — — — — — итеративное линейное программирование (Фиасконаро) 115—126
— — — — — — линейное программирование 110—115
— — — — — — функции окна 100—105
— — — — — — частотная дискретизация 105—110
— — — — (одномерных), алгоритм Паркса и Макклеллана 115
— — — — — — — Хофстеттера 115
— — — — — второй алгоритм Ремеза 115
— — — — — (*n*-мерных) 142
— рекурсивных цифровых фильтров (двумерных), математические методы оптимизации 173
— — — — — — поворот частотной характеристики одномерного фильтра 156—163
— — — — — — разделимые фильтры 156
— — — — — — сдвиг частотной характеристики двумерного фильтра нижних частот 163—169
— — — — — — синтез в пространственной области 169—173
Пропускная способность канала 227
Псевдоинверсия матриц 11, 40, 67
z-преобразование (двумерное) 84, 145

Радиолокатор бокового обзора 66
Разделимая дифференциальная ИКМ 54
— — — статистика изображения 55, 58
Разделимое линейное преобразование 37
Разделимость 31, 52
Разделимые фильтры 156
Разложение на главные компоненты 40, 58
— по сингулярным значениям (РСЗ) 11, 38—40, 67
— — — — — для кодирования изображений 47—52, 58
— — — — — — проектирования фильтров 63
— — — — — — реставрации изображений 11—15, 63, 67—73
Разрешение 209
— по Релею 231
Распознавание зрительных образов 9, 28
— образов 10
Реализация рекурсивных цифровых фильтров (двумерных), прямая 191
— — — — — — начальные условия 147
— — — — — — переменные состояния 189—191
— — — — — — получение нулевой ФЧХ методом последовательного соединения фильтров 187
— — — — — — — — — — суммирования 188
— — — — — — симметрирование 188
Резкость 225
Рекуррентная оценка 27
Рекурсивные цифровые фильтры (двумерные) 15, 145—192
— — — — обеспечение устойчивости методом двойной двумерной инверсии 173—178, 181—185
— — — — — — — — преобразования Гильберта 178—187
— — — — проверка устойчивости 16, 148—155
— — — — требования к ФЧХ 17—21
— — — — условия устойчивости для фильтра первого порядка 151
— — — — физическая реализуемость 145
— — — — шум квантования 16
Рентгеновские аппараты с прямоугольным антикатодом 66
Реставрация изображения 9, 21, 32, 63—77, 193—270, 271, 274
— — априорная информация 24, 27, 32, 194, 202, 233—240
— — вырожденные изображения 31
— — дискретная противосвертка 221—224
— — инверсная фильтрация 208—214
— — искажения вследствие линейного движения 216, 217
— — итеративные методы 21, 22, 241
— — компенсация передаточной функции 208
— — краевой эффект 15
— — метод ван Циттерта 219—221
— — — Джанссона — ван Циттерта 246—248

Реставрация изображения, метод квадратичного программирования 26, 27
— — — Монте-Карло (по Фридену) 261—264
— — — — оптимизации Бакуса — Гильберта 228—230
— — — проекций 21—26
— — — сглаживания Филлипса 201—203
— — — Шелла и Биро 241—246
— — методы математического программирования 26, 27
— — — преобразования Фурье 34
— — нелинейные методы 205, 241, 247, 265
— — непосредственная инверсия матриц 199, 200
— — неразделимое пространственно-зависимое искажение 27, 75—77
— — — пространственно-инвариантное искажение 73—75
— — обобщенный метод Туми 203—205
— — ограничение по положительности 238—240
— — — сверху 239, 240
— — ограниченная пространственная протяженность 233—240
— — оценка по максимуму правдоподобия 257
— — — — — энтропии 248—261
— — перечень неоптических приложений 195
— — подавление постоянной составляющей 215, 216
— — противосвертка с ограничениями (по Мак-Адаму) 26, 27
— — псевдоинверсия матриц 11—15, 63
— — разделимое пространственно-зависимое искажение 66—73
— — — пространственно-инвариантное искажение 64, 65
— — разложение по сингулярным значениям 11—15, 63, 67—73
— — рекуррентная оценка (фильтры Калмана — Бьюси) 27
— — сводка методов 265—267
— — случайная ФРТ 265
— — среднеквадратичный инверсный фильтр 19, 225
— — чистая фазовая компенсация 216—219
— — чистое сложение и вычитание 223
Сигнал 276
Симплексный алгоритм 96, 138
Сканер 28
— бегущего луча 271, 298
— с электроннолучевой трубкой 294, 298, 302
Сканирование с постоянной достоверностью 300
Спектрозональная классификация 275
Спектры поглощения 240, 246
Способы выделения отсчетов 287—293
Среднее по ансамблю 196
Среднеквадратичный инверсный фильтр 19, 225

Теорема Валле — Пуссена 93
— Винера — Хинчина 198
— о максимуме модуля 152
— отсчетов 197
Теория линейной аппроксимации 86—95
Тёплицева матрица 59, 60, 65
Точки с ограничениями 82, 110, 118, 121, 128, 131, 132

Улучшение изображений 9, 193—270, 271, 272
Унитарное преобразование 37
Уравновешенная система 282
Усиление контраста 272
Условие Хаара 91, 115, 140, 143
Условный индекс 71—73

Фаза преобразования Фурье 17
Фазовая компенсация 216—219
Физическая реализуемость 145
Фильтры верхних частот (двумерные) 88
— Калмана — Бьюси 27
— нижних частот (двумерные) 88, 105, 109, 112, 128—133
Фотоэлектронный умножитель 271, 287, 294
Функция знака 179
Функции, ограниченные по полосе пространства 197, 234
— — — — частот 197
— окна (двумерные) 81, 100—105
— — (одномерные), окно Кайзера 103, 105
— — — — Ханнинга 213
— — — — Хэмминга 103
— — — прямоугольное окно 101
— — — функция эллиптической дуги 213

Функции окна при проектировании фильтров 81, 100—105
— — — реставрации изображений 212—216, 233, 235
— Уолша 38
Функция зрачка 214
— рассеяния точки 34, 194
— — — изопланатическая 207
— — — инвариантная относительно сдвига 207
— — — пространственно-зависимая (ПЗФРТ) 34
— — — пространственно-инвариантная (ПИФРТ) 34
— — — случайная 265
— — — стационарная 207
Функция Rect 197
— sinc 197

Циклическая свертка 85
Цифровые фильтры (двумерные) нерекурсивные 15, 83, 81—144
— — — рекурсивные 145—192
— — — устойчивость 83, 148—155

Частотная дискретизация 81, 105—110
— характеристика, полоса пропускания 108
Частотная характеристика, ограничения по симметрии 87, 88, 107
— — переходная полоса 108
— — полоса задерживания 108
Чебышевская аппроксимация 89
— — (двумерная), теорема единственности 91
— — — — идентификации 90, 92
— — — — существования 90
— — — — чередования 92, 140
— система 91

Шум 28, 193, 271—307
— влияние на квантование 276—287
— — — точность классификации 275
— гауссов 278, 280
— двумерный 272, 273
— зернистости фотопленки 283—287
— квантования 282
— псевдослучайный 272
— тепловой 288
— фотонов 288
— фотоэмиссии 288

Экстраполяция полосы частот 230—233, 246
Эллиптические фильтры 168
Энергетический спектр 196, 274

Оглавление

ОТ ПЕРЕВОДЧИКОВ 5

ПРЕДИСЛОВИЕ 7

1. ВВЕДЕНИЕ (*Т. Хуанг*) 9
1.1. Что такое обработка изображений? 9
1.2. Структура книги и библиография общего характера 10
1.3. Двумерные преобразования 11
1.4. Двумерные цифровые фильтры 15
1.5. Улучшение и реставрация изображений 21
1.6. Технические средства цифровой обработки изображений 28
1.7. Другие области обработки изображений 28
Литература 29

2. ДВУМЕРНЫЕ ПРЕОБРАЗОВАНИЯ (*Г. Эндрюс*) 31
2.1. Введение 31
2.1.1. Решаемые задачи — кодирование, реставрация, выделение признаков 31
2.1.2. Линейные модели 33
2.2. Математическое представление 36
2.2.1. Представление изображения в ортогональном базисе . . . 36
2.2.2. Примеры представления изображений 40
2.2.3. Разделимость и пакетные операторы 52
2.3. Кодирование изображений 53
2.3.1. Пространственное кодирование изображений 53
2.3.2. Кодирование изображений посредством преобразований . . 56
2.4. Реставрация оригинала 63
2.4.1. Разделимая пространственно-инвариантная функция рассеяния точки 64
2.4.2. Разделимая пространственно-зависимая функция рассеяния точки 66
2.4.3. Неразделимая пространственно-инвариантная функция рассеяния точки 73
2.4.4. Неразделимая пространственно-зависимая функция рассеяния точки 75
2.5. Заключение 77
Литература 79

3. ДВУМЕРНЫЕ НЕРЕКУРСИВНЫЕ ФИЛЬТРЫ (*Дж. Фиасконаро*) 81
3.1. Теория 82
3.1.1. Двумерные дискретные системы 82
3.1.2. Теория аппроксимации 86
3.1.3. Линейное программирование 95

3.2. Алгоритмы 100
3.2.1. Функции окна 100
3.2.2. Частотная дискретизация 105
3.2.3. Прямое применение линейного программирования 110
3.2.4. Новый алгоритм 115
3.2.5. Примеры и сравнение методов 126
3.3. Краткое содержание и выводы 140
Литература 143

4. ДВУМЕРНАЯ РЕКУРСИВНАЯ ФИЛЬТРАЦИЯ (*Р. Рид, Дж. Шэнкс, С. Трейтел*) 145
4.1. Введение 145
4.2. Устойчивость 148
4.2.1. Теорема 1 (теорема Шэнкса) 148
4.2.2. Теорема 2 (теорема Хуанга) 152
4.3. Проектирование 155
4.3.1. Обобщение одномерных методов 156
4.3.2. Синтез в пространственной области 169
4.3.3. Обобщенный подход к аппроксимации и устойчивости . . . 173
4.4. Реализация 187
4.4.1. Фазочастотная характеристика двумерного рекурсивного фильтра 187
4.4.2. Реализация путем применения переменных состояния . . . 189
Литература 191

5. УЛУЧШЕНИЕ И РЕСТАВРАЦИЯ ИЗОБРАЖЕНИЯ (*Б. Фриден*) . . 193
5.1. Перечень неоптических приложений и библиография 195
5.2. Математическое введение, обозначения и определения 196
5.3. Интуитивные методы реставрации 199
5.3.1. Инверсия матрицы уравнения дискретного отображения . . 199
5.4. Метод сглаживания Филлипса 201
5.4.1. Вывод 201
5.4.2. Обсуждение 202
5.4.3. Достоинства и недостатки метода в применении к улучшению изображения 203
5.5. Обобщенный метод Туми 203
5.5.1. Достоинства и недостатки метода в применении к улучшению изображений 205
5.6. Последующие методы 205
5.7. Линейные или нелинейные методы? 205
5.8. Принцип метода фильтрации 207
5.9. Инверсная фильтрация 208
5.9.1. Оптимальная полоса частот обработки 211
5.9.2. Использование функций окна 212
5.10. Прочие частные линейные методы 214
5.10.1. Подавление постоянной составляющей 215
5.10.2. Устранение искажений движения путем свертки с маской 216
5.10.3. Чистая фазовая компенсация 216
5.10.4. Метод последовательной свертки ван Циттерта 219
5.10.5. Методы дискретной противосвертки 221

5.11. Линейные методы, основанные на статистическом подходе . . . 225
5.11.1. Метод фильтрации по Винеру с учетом ограничения резкости . . . 225
5.11.2. Метод оптимизации Бакуса — Гильберта . . . 228
5.12. Возможности экстраполяции полосы частот . . . 230
5.13. Центральная роль априорного знания . . . 233
5.13.1. Знание ограниченной пространственной протяженности . . 233
5.13.2. Знание положительности: влияние на паразитную осцилляцию и разрешение . . . 238
5.13.3. Знание верхней границы . . . 239
5.14. Общие замечания о методах реставрации с ограничением по положительности . . . 240
5.15. Метод Шелла и Биро . . . 241
5.15.1. Обсуждение . . . 246
5.16. Метод Джанссона — ван Циттерта . . . 246
5.17. Реставрация по максимуму энтропии: две идеи, два подхода . . 248
5.18. Решение Бурга для максимума энтропии . . . 251
5.18.1. Обсуждение . . . 252
5.19. Решение Фридена для максимума энтропии . . . 255
5.19.1. Обсуждение . . . 257
5.19.2. Контрольные реставрации . . . 257
5.20. Метод реставрации Монте-Карло . . . 261
5.21. Прочие методы . . . 264
5.22. Сводка результатов . . . 265
5.23. Возможные направления будущих исследований . . . 265
Литература . . . 268

6. ВЛИЯНИЕ ШУМА АППАРАТУРЫ ЦИФРОВОЙ ОБРАБОТКИ ИЗОБРАЖЕНИЙ (*Ф. Биллингсли*) . . . 271
6.1. Влияние шума на зрительное восприятие и анализ изображения 271
6.1.1. Зрительное восприятие изображения . . . 272
6.1.2. Влияние шума на обработку данных . . . 273
6.1.3. Дополнительные соображения . . . 276
6.2. Влияние шума на квантование . . . 276
6.2.1. Квантование сигнала с шумом . . . 277
6.2.2. Обнаружимость двух участков различной яркости . . . 280
6.2.3. Установление меры качества работы . . . 282
6.2.4. Шум системы . . . 283
6.2.5. Шум зернистости фотопленки . . . 283
6.3. Полоса частот системы и способы выделения отсчетов . . . 287
6.3.1. Шум фотодетектора . . . 287
6.3.2. Точность измерений . . . 290
6.4. Практические случаи . . . 294
6.4.1. Случай 1: постоянные параметры системы . . . 294
6.4.2. Случай 2: переменные параметры системы . . . 295
6.5. Выводы . . . 305
Литература . . . 306

БИБЛИОГРАФИЯ . . . 308
ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ . . . 311

УВАЖАЕМЫЙ ЧИТАТЕЛЬ!

Ваши замечания о содержании книги, ее оформлении, качестве перевода и другие просим посылать по адресу: 129820, Москва, И-110, ГСП, 1-й Рижский пер., 2, изд-во «Мир».

ОБРАБОТКА ИЗОБРАЖЕНИЙ
И ЦИФРОВАЯ ФИЛЬТРАЦИЯ

Под редакцией Т. Хуанга

Ст. научный редактор Телесиин Н. Л.
Мл. научные редакторы Малейиова Л. Н., Цлаф В. Н.
Художиик В. А. Сысоев
Художественный редактор Л. Б. Безручеиков
Технический редактор Т. А. Максимова
Корректор Т. П. Пашковская

ИБ № 1043

Сдано в иабор 28.09.78. Подписано к печати 24.05.79. Формат 60 × 90$^1/_{16}$. Бумага типографская № 1. Гарнитура латинская. Изд. № 2/9149. Печать высокая. 10 бум. л., 20 усл. печ. л., 18,99 уч.-изд. л. Тираж 4800 экз. Зак. 1311. Цена 2 р. 70 к.

Издательство «Мир», Москва, 1-й Рижский пер., д. 2.

Ордена Трудового Красного Знамени Ленинградская типография № 2 имеии Евгении Соколовой «Союзполиграфпрома» при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 198052, Ленинград, Л-52, Измайловский проспект, 29

ОПЕЧАТКА
К КНИГЕ Т. ХУАНГА «ОБРАБОТКА ИЗОБРАЖЕНИЙ И ЦИФРОВАЯ ФИЛЬТРАЦИЯ»

стр. 252

Напечатано	Должно быть	По чьей вине
5.181. Осуждение	5.18.1. Обсуждение	Типография

Зак. 1311

Обработка изображений
и цифровая фильтрация